On-the-fly Dense 3D Surface Reconstruction for Geometry-Aware Augmented Reality

by
Long Chen

Faculty of Science & Technology

Bournemouth University

Supervised by Prof. Wen Tang, Prof. Nigel W. John
and Prof. Jian J. Zhang

A thesis submitted in partial fulfilment of the
requirements of Bournemouth University for the degree of
Doctor of Philosophy

Oct. 2018

Copyright Statement

Abstract

Augmented Reality (AR) is an emerging technology that makes seamless connections between virtual space and the real world by superimposing computer-generated information onto the real-world environment. AR can provide additional information in a more intuitive and natural way than any other information-delivery method that a human has ever invented. Camera tracking is the enabling technology for AR and has been well studied for the last few decades. Apart from the tracking problems, sensing and perception of the surrounding environment are also very important and challenging problems. Although there are existing hardware solutions such as Microsoft Kinect and HoloLens that can sense and build the environmental structure, they are either too bulky or too expensive for AR.

In this thesis, the challenging real-time dense 3D surface reconstruction technologies are studied and reformulated for the reinvention of basic position-aware AR towards geometry-aware and the outlook of context-aware AR. We initially propose to reconstruct the dense environmental surface using the sparse point from Simultaneous Localisation and Mapping (SLAM), but this approach is prone to fail in challenging Minimally Invasive Surgery (MIS) scenes such as the presence of deformation and surgical smoke. We subsequently adopt stereo vision with SLAM for more accurate and robust results. With the success of deep learning technology in recent years, we present learning based single image reconstruction and achieve the state-of-the-art results. Moreover, we proposed context-aware AR, one step further from purely geometry-aware

AR towards the high-level conceptual interaction modelling in complex AR environment for enhanced user experience. Finally, a learning-based smoke removal method is proposed to ensure an accurate and robust reconstruction under extreme conditions such as the presence of surgical smoke.

Contents

List of Figures

Acknowledgements

This work would not have been possible without the advice and support of my supervisory team and family. First and foremost, I would like to express my sincere gratitude to my first supervisor Professor Wen Tang for her enduring support and inspiration. I have been extremely lucky to have her who was always reachable and cared so much about my work. I appreciate all the fruitful discussions we had that generated lots of ideas in this thesis. My sincere thanks also goes to my second supervisor Professor Nigel W. John from the University of Chester, who provided the funding jointly with Bournemouth University that allowed me to undertake this research, and also provided me with many helpful comments and inspirations. I am very thankful to my third supervisor Professor Jian Jun Zhang, who gave me many useful suggestions about the direction of my research.

I must express my gratitude to Ziyuan, my wife, who has been extremely supportive of me throughout this entire process. I thank my parents for providing unconditional love and support throughout my educational career, encouraging me in all of my pursuits and inspiring me to follow my dreams.

It was a great pleasure to do a PhD in such a beautiful coastal resort town, with beautiful beaches, mountains, forests and friendly people. Probably after few years leaving Bournemouth, I will always look back fondly at these past few years of researching on fascinating problems with great people in a beautiful place by the sea.

Author's Declaration

I, Long Chen, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Acronyms

AR Augmented Reality.

BA Bundle Adjustment.

BCI Brain-Computer Interfaces.

BM Block Matching.

BOW Bags of Words.

BRIEF Binary Robust Independent Elementary Features.

BRISK Binary Robust Invariant Scalable Keypoints.

CBFM Constraint-Based Factorization Methods.

CNN Convolutional Neural Networks.

CRF Conditional Random Fields.

CT Computed Tomography.

DaG Detection-after-Generation.

DCNN Deep Convolutional Neural Networks.

DCP Dark-Channel Prior.

EKF Extended Kalman Filter.

FCN Fully Convolutional Network.

FEM Finite Element Model.

FOV Field Of View.

FREAL Fast Retina Keypoint.

GAN Generative Adversarial Network.

GCN Generative Collaborative Networks.

GPU Graphics Processing Unit.

HAR Haptic Augmented Reality.

HMA Hierarchical Multi-Affine.

HMD Head Mounted Display.

HOG Histograms of Gradients.

ICP Iterative Closest Point.

IMU Inertial Measurement Unit.

IU Intersection over Union.

IV Integral Videography.

KLT Kanade-Lucas Tomasi.

LDA Latent Dirichlet Allocation.

MINC Materials in Context Database.

MIS Minimally Invasive Surgery.

MLS Moving Least Square.

MR Mixed Reality.

MRF Markov Random Field.

MRI Magnetic Resonance Imaging.

ORB Oriented FAST and Rotated BRIEF.

PCA Principle Component Analysis.

PM Patch Matching.

PSNR Peak Signal-to-Noise Ratio.

PTAM Parallel Tracking and Mapping.

RANSAC Random Sample Consensus.

RMSD Root Mean Square Distance.

RNN Recurrent Neural Networks.

SAR Spatial Augmented Reality.

SFS Shape from Shading.

SIFT Scale Invariant Feature Transform.

SLAM Simultaneous Localisation and Mapping.

SSIM Structural Similarity Index.

SURF Speed-up Robust Feature.

VBEM Variational Bayes Expectation Maximization.

VR Virtual Reality.

ZNCC Zero-Mean Normalized Cross Correlation.

Chapter 1

Introduction

1.1 Background

Augmented Reality (AR) technology makes the use of computer-generated virtual information within the real world to enhance the human perception of the world. Azuma (Azuma 1997) defines AR as a combination of the virtual world and the real world, real-time interactions and three-dimensional registration. AR will soon be ubiquitous in many application areas ranging from personal information systems, industrial and military simulations, office use, digital games, to education and training (Van Krevelen and Poelman 2010).

Since AR was first introduced in mid-1990's (Milgram and Kishino 1994), AR technology in medicine was among the six established potential areas of AR applications (Azuma 1997). However, due to the lack of enabling technologies such as real-time tracking and display, the use of AR was still hypothetical in that time. The potential impact of such technology on professional practices has been envisaged, from using AR to support not only medical education and training, but also to help in surgical planning and to guide complex procedures. One prominent example is the possibility of using AR to visualize data from medical imaging scanners (e.g., MRI and CT) during Minimally Invasive Surgery (MIS) (Fuchs et al. 1998). By directly linking patient data such as 3D anatomical models with complex surgical scenes, AR can fuse real scenes with virtual anatomical models, thereby offering a

rich source of information to guide intrinsic movements for surgeons.

Given the technology advances in the recent years, the potential of AR has not yet been unleashed, due partly to technology limitations and costs, but also to the lack of systematic approach and understandings of this cross-cutting and inherently multidisciplinary field. Since 2016, portable high-performance augmented reality head-mounted display platforms became available to consumers such as Microsoft's HoloLens and Magic Leap's display. At the same time, new tracking algorithms with improved performance, new calibration methods and real-time interactions in AR have also emerged. Despite continued hardware and software development has stimulated a surge in mixed and augmented reality research projects (Zhou et al. 2008) (Van Krevelen and Poelman 2010), many challenging problems must be overcome before the use of AR can be a commonplace in everyday life. A revolutionary impact of AR, however, is yet to be achieved.

1.2 Main Challenges

Real-time performance and minimum latency are pre-requisites in most medical applications. A typical AR system consists of multiple modules working together (image capture module, image detection and tracking module, content rendering module, image fusion module, display module, etc.) each of which has its own computational demands and each component can contribute to latency. Especially in medical related applications, the use of high-fidelity patient specific data (such as the offline models reconstructed from pre-operative CT/MRI and the online models from intra-operative MRI/X-rays) is time-consuming in reconstruction, storing and retrieving, and displaying processes. Although improvements to hardware and software continue to address these technology problems. New generation hardware devices such as the Hololens, Magic Leap, and Meta Glasses will soon enable the real-time AR in medical practices. New algorithms and software based on graphics processing unit (GPU) will further shorten the system latencies while keeping high precision. According to (Lambert et al. 2016), the frame rate of the video should be

as high as 100 FPS at least to enable the doctor to detect minor changes. Whereas, the AR system based on advanced tracking technologies such as Simultaneous Localisation and Mapping (SLAM) system currently can only reach quasi real-time performance around 20 FPS (Mur-Artal et al. 2015).

High precision is another challenge must be stressed in medical applications. Many clinical procedures could benefit greatly from AR if patient specific data (such as anatomic structures, vessels and tumour locations etc.) can be accurately delivered to the clinician to help guide the procedure. In such a system, however, if the augmented content were superimposed in the wrong position then the clinician could be misled and cause a serious medical accident. Many researchers are obtaining accuracy to within a few millimetres and this may be sufficient for some procedures and applications (an anatomy education tool, for example). Other procedures will need sub-millimetre accuracy. Automatic setup and calibration then become more critical. It remains challenging to find a balance between speed and accuracy as they are both very important in medical applications of AR.

Apart from the tracking problems above, sensing and perception of the surrounding environment are also great challenges that limit the future development of AR. Without the understanding of the environment such as the structural information, depth and material, it is like a blind person entering into an unknown room and does not know what to sit on. For example, when using the very basic AR application – virtual object placement, the system wouldn't know where to place the objects in the virtual world, as the location might be occupied by a real world object. This might lead to an awkward situation where the virtual objects are placed in the wall and ruin the AR user experience. Therefore, it is essential for the AR system to have a knowledge of the surrounding environment and enable geometry-aware AR.

Aside from real-world AR scenarios, in medicine, sensing and perception of the surrounding environment are also very important. Offline high-fidelity image capture and 3D reconstruction from medical imaging devices such as CT, MR can provide some of the patient specific data needed for AR surgical guidance, but real-time high-fidelity online model reconstruction is also needed, for example, to handle tissue

deformation. There is an urgent need for more research into real time high-fidelity 3D model reconstruction using online modalities such as video from endoscopic cameras. This will reduce the disparity between offline and real-time reconstruction performance capture.

1.3 Research Aims and Contributions

This research aims to investigate real-time 3D surface reconstruction technologies using different modalities of camera (mono/stereo) for geometry-aware AR, specifically for both surgical guidance and real-world games. During the development of this PhD project, deep learning based methods have been deployed to solve single-image reconstruction, semantic reconstruction and also smoke removal to improve the robustness of surface reconstruction.

More specifically, the main contributions of this work are:

- Hierarchical classifying and reviewing the latest trend and future development of Augmented Reality using an automatical data-mining approach *(Chapter 2)*.
- Revisiting and exploiting the SLAM system to be used for tracking and dense reconstruction for monocular camera geometry-aware AR *(Chapter 4)*.
- Integrating and fusing the information provided by a stereo camera for more accurate on-the-fly global surface dense reconstruction *(Chapter 5)*.
- Exploring and employing the latest deep learning based method for single image depth estimation and 3D reconstruction *(Chapter 6)*.
- Initiating a higher level Context-Aware AR system beyond the current Geometry-Aware AR for advanced virtual-real interaction *(Chapter 7)*.
- Enhancing the tracking and reconstruction robustness by a novel learning-based smoke removal framework *(Chapter 8)*.

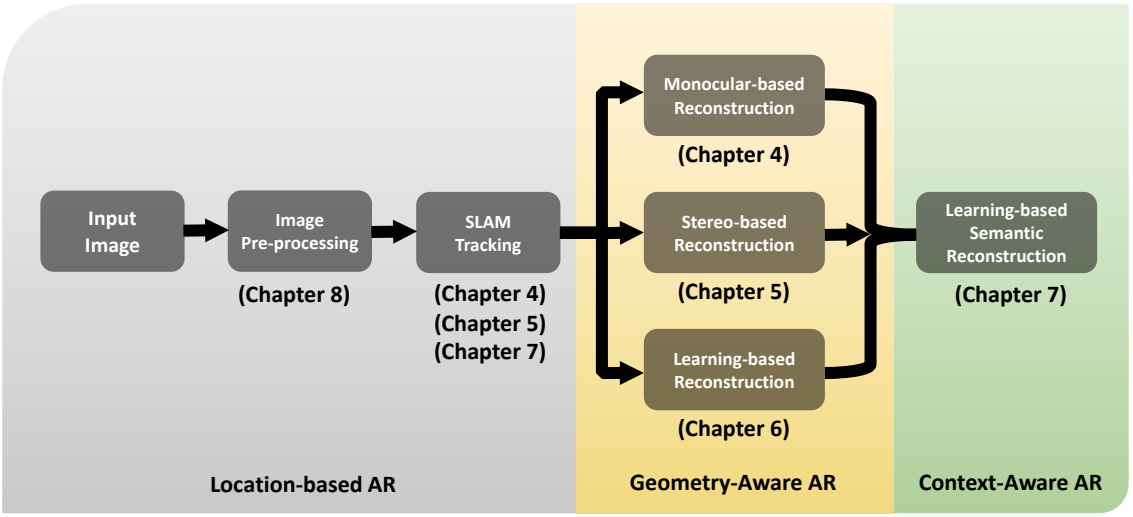


Figure 1.1: The "Big Picture" of this thesis.

In Figure 1.1, the big picture of this thesis is summarized. From a high-level perspective, Chapter 4, 5, 7, 8 covers the knowledge of SLAM tracking and image learning-based image pre-processing for the stage of Location based AR. In Chapter 4, 5 and 6, the proposed three different surface reconstruction methods are combined with SLAM tracking for the goal of Geometry-Aware AR. In Chapter 7, we adopt learning-based semantic reconstruction for the conception of high-level Context-Aware AR.

1.4 Structure of the Following Chapters

- **Chapter 2 - Research Topic Classification, Trend Analysis and Technology Review:** In this chapter, a classification of AR has been obtained by applying an unbiased text mining method to a database of 1,403 relevant research papers published over the last two decades. The classification results reveal a taxonomy for the development of AR research during this period as well as suggesting future trends. Then the classification results are used to analyse the technology and applications developed in the last five years. The objective of this chapter is to aid researchers to focus on the areas where tech-

nology advancements in AR are most needed, as well as providing medical practitioners with a useful source of reference. This chapter is presented and published in the proceedings of the 16th IEEE International Symposium on Mixed and Augmented Reality (ISMAR) 2017 (Chen et al. 2017a).

- **Chapter 3 - Problem Statement & Literature Review:** In this chapter, the research problem is identified and the research hypothesis is given. Based on the research problem and hypothesis, a detailed literature review is given on the topic of the latest AR camera tracking and surface reconstruction literature.

- **Chapter 4 - Monocular-based Online Dense Surface Reconstruction for GA-AR:** In this chapter, the concept of Geometry-Aware AR (GA-AR) is first introduced. A novel intra-operative dense surface reconstruction framework is then proposed that is capable of providing geometry information from monocular Minimally Invasive Surgery videos for GA-AR applications such as surgical guidance, intra-operative measurements and providing mesh-based depth cues. This chapter is published in the journal of Computer Methods and Programs in Biomedicine, Volume 158, May 2018, Pages 135-146 (Chen et al. 2018b).

- **Chapter 5 - Stereo-based Online Global Surface Reconstruction for GA-AR:** In this chapter, a novel stereo-based real-time AR framework is proposed that provides 3D geometric information for accurate AR content registration and overlay in MIS. We propose a new approach to achieving robust 3D tracking through a feature-based SLAM for real-time performance and accuracy required for endoscopy camera tracking. To obtain accurate geometric information, we incrementally build a dense 3D point cloud by using Zero-Mean Normalized Cross Correlation (ZNCC) stereo matching. Therefore, our framework handles the challenging situations of rapid endoscopy movements with robust real-time tracking, while providing an interactive geometry-aware AR environment. This chapter is presented in the 11th MICCAI workshop on

Augmented Environments for Computer-Assisted Interventions (AECAI) and published in the journal of Healthcare Technology Letters, Volume 4, Issue 5, Oct 2017, Pages 163–167 (Chen et al. 2017c).

- **Chapter 6 - Learning-based Monocular Image Depth Estimation and 3D Reconstruction:** In this chapter, a novel framework for depth estimation and reconstruction from monocular images is presented with the bonus of estimating corresponding confidence in a self-supervised manner. A fully differential patch-based cost function is proposed by using the Zero-Mean Normalized Cross Correlation (ZNCC) that takes multi-scale patches as a matching strategy. This approach greatly increases the accuracy and robustness of the depth learning. In addition, the proposed patch-based cost function can provide a 0 to 1 confidence, which is then used to supervise the training of a parallel network for confidence map learning and estimation. Evaluation on public dataset shows that our method outperforms the state-of-the-art results. The content of this chapter is currently under review by the Neurocomputing journal.

- **Chapter 7 - From Geometry-Aware AR to Context-Aware AR:** This chapter is an initiative beyond the current Geometry-Aware AR approach towards the higher level Context-Aware AR. An interactive MR framework is proposed based on the latest SLAM technology and deep learning based material recognition for providing a whole new AR experience with context-wise interaction. Quantitative and qualitative evaluations were carried out and described in this Chapter. The results show that the framework delivers accurate and fast semantic information in interactive AR environment, providing effective context level interactions. The content of this chapter is presented in the 16th IEEE International Symposium on Mixed and Augmented Reality (ISMAR) 2017 (Chen et al. 2017b) and is being reviewed by the Computer Graphics Forum journal.

- **Chapter 8 - Increase Tracking and Reconstruction Robustness –**

Learning-based Image Smoke Removal: This chapter includes an additional work to the main topic of this thesis – increasing tracking and reconstruction robustness under extreme conditions e.g. under heavy surgical smoke. A new unsupervised learning framework for high quality pixel-wise smoke detection and removal is presented in this chapter. A novel generative-collaborative learning scheme is proposed that decomposes the de-smoke task into two separate tasks: smoke detection and smoke removal. While using the detection network as prior knowledge, it is also used as a loss function to maximize its support for the removal of network training. The quantitative and qualitative study shows that our training framework outperforms the latest GAN framework (such as PIX2PIX) and the state-of-the-art de-smoking approaches. The content of this chapter is presented in 2018 Hamlyn Symposium on Medical Robotics (Chen et al. 2018a) and under review by the journal of IEEE Transactions on Medical Imaging.

1.5 List of Publications

The following publications are a direct result of the research carried out in this thesis:

Conference Papers

Chen, L. and Tang, W., 2016. MathRun: an adaptive mental arithmetic game usinga quantitative performance model. *30th International BCS Human Computer Interaction Conference* (HCI 2016).

Chen, L., Day, T. W., Tang, W. and John, N. W., 2017a. Recent developments and future challenges in medical mixed reality. *2017 IEEE International Symposiumon Mixed and Augmented Reality (ISMAR)*, IEEE.

Chen, L., Francis, K. and Tang, W., 2017b. Semantic augmented reality environment with material-aware physical interactions. *2017 IEEE International Symposiumon Mixed and Augmented Reality (ISMAR-Adjunct)*, IEEE

Chen, L., Tang, W. and John, N. W., 2017c. Real-time geometry-aware aug-

mented reality in minimally invasive surgery. *2017 MICCAI workshop on Augmented Environments for Computer-Assisted Interventions (AECAI).*

Chen, L., Tang, W. and John, N. W., 2018. Unsupervised Learning of Surgical Smoke Removal from Simulation. *2018 Hamlyn Symposium on Medical Robotics (HSMR).*

Gao, Q. H., Wan, T. R., Tang, W. and **Chen, L.**, 2017a. A stable and accurate marker-less augmented reality registration method. *2017 International Conferenceon Cyberworlds (CW)*, IEEE.

Gao, Q. H., Wan, T. R., Tang, W., **Chen, L.** and Zhang, K. B., 2017b. An improved augmented reality registration method based on visual SLAM. *E-Learning and Games, Springer International Publishing, 11–19.*

Journal Papers

Chen, L., Tang, W. and John, N. W., 2017c. Real-time geometry-aware augmented reality in minimally invasive surgery. *Healthcare Technology Letters*, 4 (5), 163–167.

Chen, L., Tang, W., John, N. W., Wan, T. R. and Zhang, J. J., 2018. SLAM-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality. *Computer Methods and Programs in Biomedicine*, 158, 135–146

Papers Currently Under Review

Chen, L., Tang, W., John, N. W., Wan, T. R. and Zhang, J. J. Context-Aware Mixed Reality: A Learning-based Framework for Semantic-level Interaction.

Chen, L., Tang, W., John, N. W., Wan, T. R. and Zhang, J. J. De-smokeGCN: Generative Cooperative Networks for Joint Surgical Smoke Detection and Removal.

Chen, L., Tang, W., John, N. W., Wan, T. R. Self-Supervised Monocular Image Depth Learning and Confidence Estimation.

Chapter 2

Research Topic Classification, Trend Analysis and Technology Review

In this Chapter, we perform a through classification and review of AR technologies in medicine. Medical AR is an very broad topic that can be viewed as a multidimensional domain (see Fig. 2.1) with a crossover of many technologies (e.g. camera tracking, visual displays, computer graphics, robotic vision, and computer vision etc.) and applications (e.g. medical training, rehabilitation, intra-operative navigation, guided surgery). In this research, we discard the traditional literature review method and present a novel classification method and by combining text mining, topic generation/clustering, and taxonomic review for a better understanding of development trends, current issues and future directions.

2.1 Automatic Classification of AR using Data-Mining

Bibliometric methods are the most common approaches used in identifying research trends by analysing scientific publications(Li et al. 2009) (Hung and Zhang 2012)

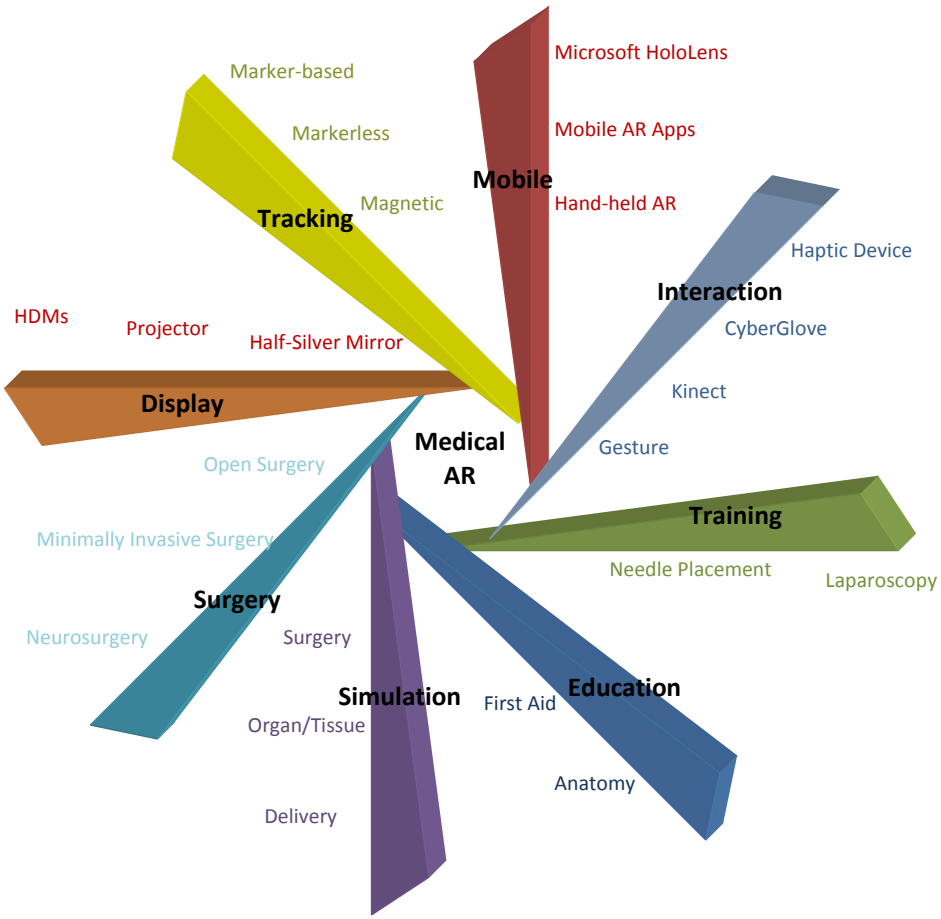


Figure 2.1: Many different technologies and applications fall within the medical AR domain.

(Vanga et al. 2015) (Dey et al. 2016). These methods typically make predictions by measuring certain indicators such as geographical distributions of research institutes and the annual growth of publications, as well as citation counts(Gireesh et al. 2008). Usually a manual classification process is carried out (Dey et al. 2016), which is inefficient and also can be affected by personal experience. Especially in a large domain such as Medical AR, it is challenging to identify the classification and review the trend. Our approach is to analyse the relevant related papers retrieved from different periods, whilst introducing a novel method to automatically decompose the overarching topic (medical mixed reality) into relevant sub-topics that can be analysed separately. We make use of a generative probabilistic model for text mining – Latent Dirichlet Allocation (LDA) (Blei et al. 2003) to automatically classify and

generate the categories, achieving an unbiased review process.

2.1.1 Data Source

The source database used in this analysis is Scopus, which contains the largest abstract and citation database of peer-reviewed literature obtained from more than 5,000 international publishers(Elsevier 2018). Scopus contains articles published after 1995 (Reuters 2016), therefore, encompassing the main period of growth in AR, and helps both in keyword searching and citation analysis.

2.1.2 Selection Criteria

The regular expression "(mixed OR augmented) reality medic*" is used to retrieve all articles related to augmented reality and mixed reality in medicine, capturing "augmented reality medicine", "augmented reality medical", and "mixed reality medicine", "mixed reality medical". A total of 1,425 articles were retrieved within the 21 year period from 1995 to 2015, of which 1,403 abstracts were accessed. We initially categorised these articles into seven chronological periods, one for every three years. Abstracts of these articles are then used to generate topics and for trend analysis, as they provide more comprehensive information about an article than its title and keywords alone. The whole selection process is carried out automatically with no manual intervention.

2.1.3 Text Mining

To identify the key topics discussed in a large number of articles, we employ the Latent Dirichlet Allocation (LDA)(Blei et al. 2003) method to automatically interpret and cluster words and documents into different topics. This text mining method has been widely used in recommendation systems such as web search engines and advertising applications. LDA is a generative probabilistic model of a corpus. It regards documents (d) as random mixtures over latent topics (t), $p(t|d)$, where every topic is characterized by a distribution over words (w), $p(w|t)$. The method uses

Table 2.1: Topic Clustering Results from the LDA Model

Topic1 **"Treatment"**		*Topic 2* **"Education"**		*Topic 3* **"Rehabilitation"**		*Topic 4* **"Surgery"**		*Topic 5* **"Training"**	
term	**weight**	**term**	**weight**	**term**	**weight**	**term**	**weight**	**term**	**weight**
treatment	0.007494	learning	0.01602	physical	0.01383	surgical	0.05450	training	0.03044
clinical	0.007309	development	0.00877	rehabilitation	0.01147	surgery	0.02764	performance	0.01361
primary	0.004333	education	0.00854	environment	0.01112	surgeon	0.01176	laparoscopic	0.01332
qualitative	0.003793	potential	0.00812	game	0.00837	invasive	0.01167	skills	0.01208
focus	0.004165	different	0.00793	therapy	0.00729	minimally	0.01148	simulator	0.01198

Topic6 **"Interaction"**		*Topic 7* **"Mobile"**		*Topic 8* **"Display"**		*Topic 9* **"Registration"**		*Topic 10* **"Tracking"**	
term	**weight**	**term**	**weight**	**term**	**weight**	**term**	**weight**	**term**	**weight**
human	0.019901	software	0.01684	visualization	0.03260	registration	0.01417	tracking	0.02418
interaction	0.014849	mobile	0.01556	data	0.03177	segmentation	0.00942	accuracy	0.01584
haptic	0.01439	support	0.00905	display	0.00984	accurate	0.00765	camera	0.01454
feedback	0.013308	online	0.00874	navigation	0.01278	deformation	0.00762	target	0.01347
interface	0.009382	social	0.00835	planning	0.01225	motion	0.00754	registration	0.01186

the following formula:

$$p(w|d) = p(w|t) * p(t|d) \tag{2.1}$$

where $p(w|d)$ represents the probability of a certain word in a certain document under a certain topic. Word-topic distribution $p(w|t)$ and topic-document distribution $p(t|d)$ are randomly selected and then LDA iteratively updates and estimates probabilities until the system convergences. As a result, we identify the relationships amongst documents, topics and words. We input all of the downloaded abstracts into the LDA model and tested a range of parameters to use with it. We empirically derived the value of ten topics as optimal for the best encapsulation of the field.

2.1.4 Topic Generation

Table 2.1 summarizes the output showing the ten topics identified with the associated term and weight distributions after convergence. We manually assign one word (shown in quotation marks) that was the best representation of each topic. The general methodology uses the weighting as the primary selection parameter but also takes into account the descriptive keyword for that topic. Topics 1, 5, 9 and 10 just use the keyword with the highest weighting. For Topic 2, although "education" did

not have the highest weighting, we consider it to be a more representative term. For Topic 3, "physical" is a sub category of "rehabilitation" and so we use the latter as the more generic term. In Topic 4, "surgical" and "surgery" are fundamentally the same. For Topic 6, "interaction" is the most representative keyword, and the same principle applies to Topics 7 and 8.

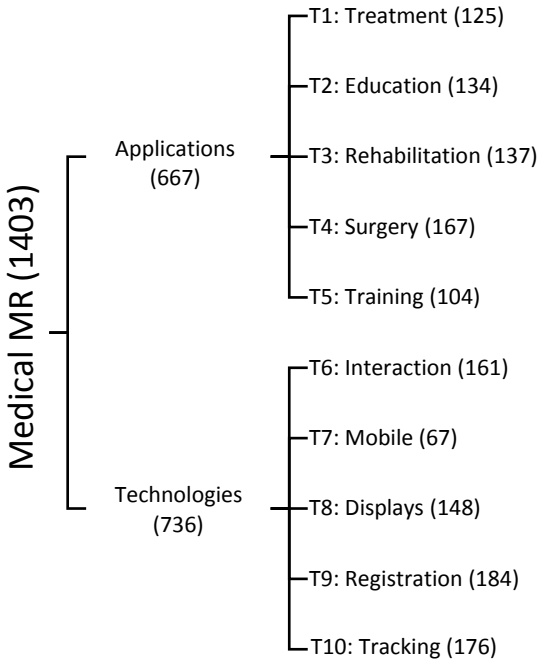


Figure 2.2: Hierarchical taxonomy of Medical AR generated by the LDA model.

Figure 2.2 represents a hierarchical taxonomy of the results. The overarching "Medical AR" topic with 1,403 articles has been divided into two main subcategories: applications and technologies, with 667 and 736 articles respectively. Within applications, the surgical topic has the largest number of articles (167), followed by rehabilitation (137), education (134), treatment (125) and training (104). Within technologies, registration is the most discussed topic (184 articles), followed by tracking (176), interaction (161), displays (148) and mobile technologies (67).

2.2 AR Trend Analysis

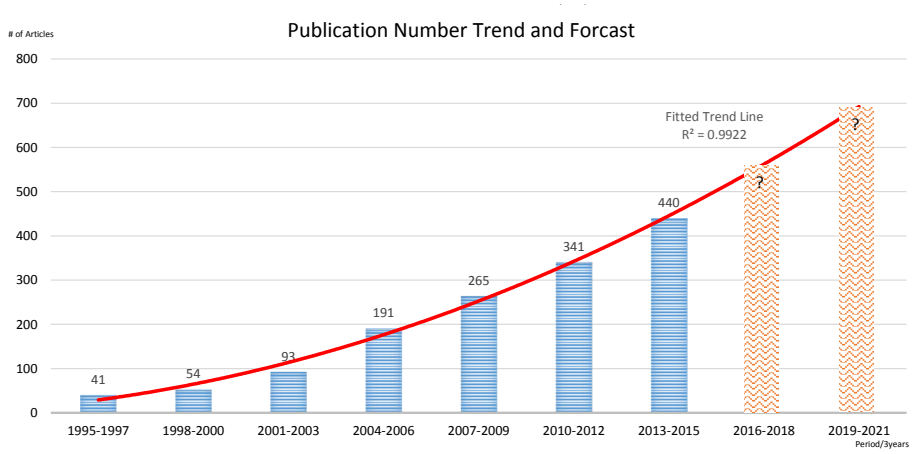
Each of the ten topics of interest identified by the LDA model has a list of articles associated with them. An article matrix was constructed based on the topic and the attributes for the seven chronological periods being analysed. Figure 2.3 summarizes the trends identified subdivided into three year periods (1995-97, 1998-2000, 2001-03, 2004-06, 2007-09, 2010-12, and 2013-15). Figure 2.3(a) plots the total number of publications over the seven periods. The number of publications related to AR in medicine has increased more than 100 times from only 41 publications in 1995-1997 to 440 publications in 2013-2015. In the early 21st century (periods 2001-2003 and 2004-2006), the number of publications of AR in medicine more than doubled from 93 to 191, coinciding with the rapid development of many enabling technologies such as marker-based tracking techniques(Kato and Billinghurst 1999) and advances in Head Mounted Display (HMD) technology (Rolland and Fuchs 2000) and mobile AR devices (Olsson and Salo 2011).

Based on the observed growth pattern between 1995 and 2015, a trend line has been produced using a quadratic polynomial with a high coefficient of determination ($R^2 = 0.9935$). Extrapolating the trend line forecasts that in the three year periods (2016 to 2018, and 2019 to 2021), the number of scientific papers on the topic of AR in medicine will be accelerated, reaching around 550 and 700, respectively. The following section looks in more detail at the topic trends and then we analyse the research trends in each area.

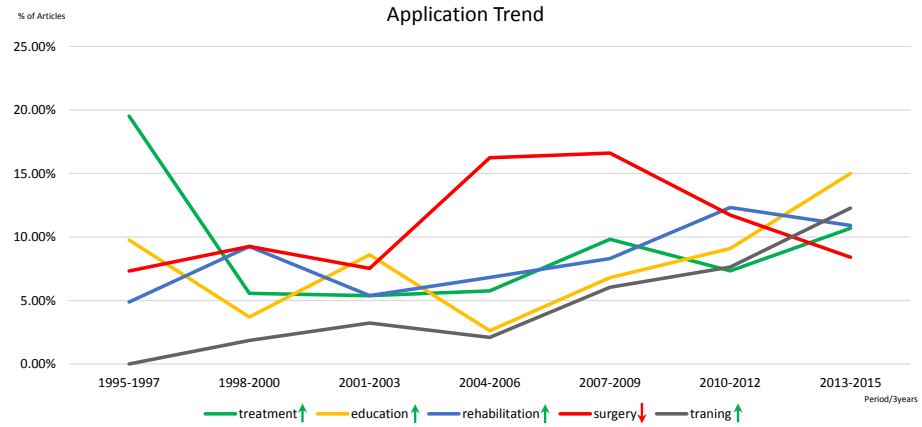
2.2.1 Applications Trends

There are a growing number of medical application areas exploring the use of AR. Fig. 2.3(b) plots the percentage of articles published for the five most popular application categories: *patient treatment, medical and patient education, rehabilitation, surgery,* and *procedures training*:

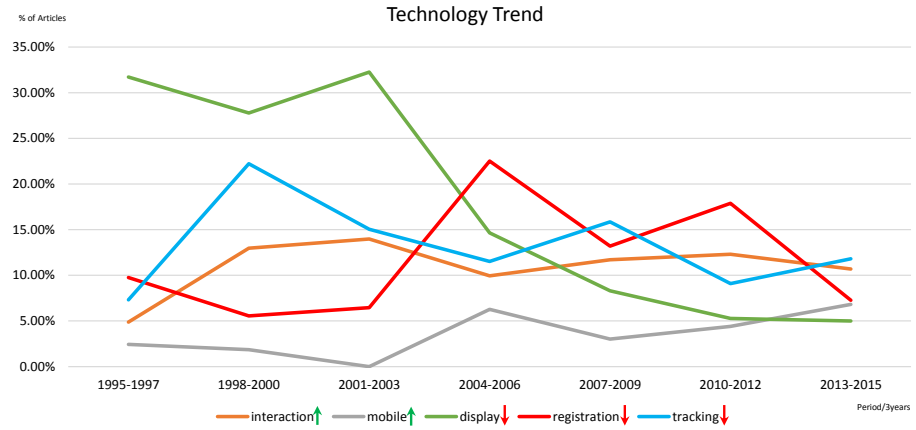
- Patient treatment was the most targeted application of AR in the earlier period with almost 20% of published articles. It remains a constant topic of interest



(a)



(b)



(c)

Figure 2.3: Trend analysis: (a) Publication Trends. (b) Application Trends. (c) Technology Trends.

with around 10% of articles continuing to investigate this topic. The fall in percentage is mostly due to the parallel rise in interest in the other medical AR categories. Education and rehabilitation topics have both fluctuated but remain close to 10% of articles published.

- A surge of interest in surgical applications can be observed between 2004 and 2009 when 16% of all articles published on medical AR addressed this topic. However, the comparative dominance of surgical applications has dropped off as activity within other categories has increased.

- Training applications in medical AR first emerged between 1998-2000. Interest in this topic has grown steadily and is also now at a similar level of interest as the other topics. Together with education, continuation of the current trends suggest that these two topics will be the most popular in the next few years. These are areas where there is no direct involvement with patients and so ethical approval may be easier to gain.

2.2.2 Technologies Trends

Within the ten topics generated by the LDA model, five key technologies have been identified: *interaction, mobile, display, registration* and *tracking* (the percentage of articles that refer to these technologies is plotted in Fig. 2.3(c)):

- Real time interaction is a crucial component of any AR application in medicine especially when interactions with patients are involved. The percentage of articles that discuss interaction in the context of medical AR increased steadily from 5% in 1995-1997 to 10% in 2013-2015.

- The use of mobile technologies is an emerging trend, which has been increased from 0% to 7% of articles across the seven periods. The main surge so far was from 2004-2006, when the advances of micro-electronics technology first enabled mobile devices to run fast enough to support AR applications. The use of mobile technologies has been more or less constant from that point

onwards. Smartphones and tablets can significantly increase the mobility and user experience as you are not tethered to a computer, or even in one fixed place as was the case with Sutherland's (Sutherland 1968) first AR prototype in the 1960s.

- Innovations in the use of display technologies was the most discussed AR topic in the early part of the time-line. However, there has been a subsequent dramatic drop in such articles, falling from 33% of articles to only 5%. This may indicate the maturity of the display technologies currently in use. The Microsoft Hololens and other new devices are expected to disrupt this trend, however.

- Tracking and registration are important enabling components for AR. They directly impact on the usability and performance of the system. These areas continue to be explored and are yet to be mature enough for complex scenarios, as reflected by the steady percentage (around 10%) of articles on tracking and registration technology from 1995 to 2015.

In the next section we summarise the latest research in medical AR using the classification scheme identified above. We restrict our analysis to publications in the last five years, citing recent survey papers wherever possible.

2.3 Review of Enabling Technologies for AR

Having identified five technology areas as listed in table 1: Interaction (topic 6), Mobile (topic 7), Display (topic 8), Registration (topic 9), and Tracking (topic 10). In this section, we provide a detailed overview of each topic area with medical applications, including developments in hardware and software algorithms.

2.3.1 Interaction

Interaction within augmented reality needs to be a natural seamless integration of the real and augmented environments. AR interactions are supported through a

variety of interaction devices (including the use of haptics), gesture based interfaces and other novel approaches that have been used in a medical AR context.

2.3.1.1 Gesture-based Interfaces

Perhaps the most natural interaction is to use our own hands to interact with virtual objects in the augmented environment. Gesture-based interactions requires the tracking of hand movement including finger movements to manipulate virtual objects. There are a number of approaches to achieve such tracking via AR markers, gloves, hand held devices, or direct optical tracking of the user's hands. For example:

- Boonbrahm and Kaewrat (Boonbrahm and Kaewrat 2014) modified AR markers to fit on fingertips and assigned the corresponding virtual fingers with physical properties such as friction, density, surface, volume and collision detection. In this way, users could interact with a virtual object with their own hands such as grasping and lifting.

- FIGI (floating interface for gesture-based interaction) (De Marsico et al. 2014) uses a wireless instrumented glove (5DT Data Glove 14 ultra) to capture finger movement and identify gestures to perform zoom and rotation tasks, select 3D medical images, and even typing on a floating virtual keyboard in mixed reality environment.

- Hochreiter *et al* (Hochreiter et al. 2015) used infra red (IR) light to detect touch events on a shell of a human head onto which a facial animation is projected from a pico projector. This physical-virtual head has been used for hands-on healthcare training, employing touch gestures such as pulling the lips apart to inspect the gums.

- A recent commercial device that uses optical tracking of the hands without the need for the user to wear gloves or markers is the Leap Motion - and several medical related applications have been reported, e.g. (Sousa Silva and Formico Rodrigues 2015).

- The Nintendo Wiimote is one example of a hand held device that has been used in medical visualization to capture gestures to interact with virtual objects (Lin et al. 2012), which the authors claim provides a more realistic interaction experience than traditional controllers such as mouse and keyboard.

2.3.1.2 Haptic Devices

The sense of touch provides important cues in most medical procedures. In AR, the physical world is still available and can be touched but any virtual object may need to make use of a haptics device to provide tactile or force feedback. A comprehensive review of the state-of-the-art haptic devices in medical training simulators before 2010 has been presented in (Coles et al. 2011a). However, there are only a few examples of these haptic devices been used in AR to date.

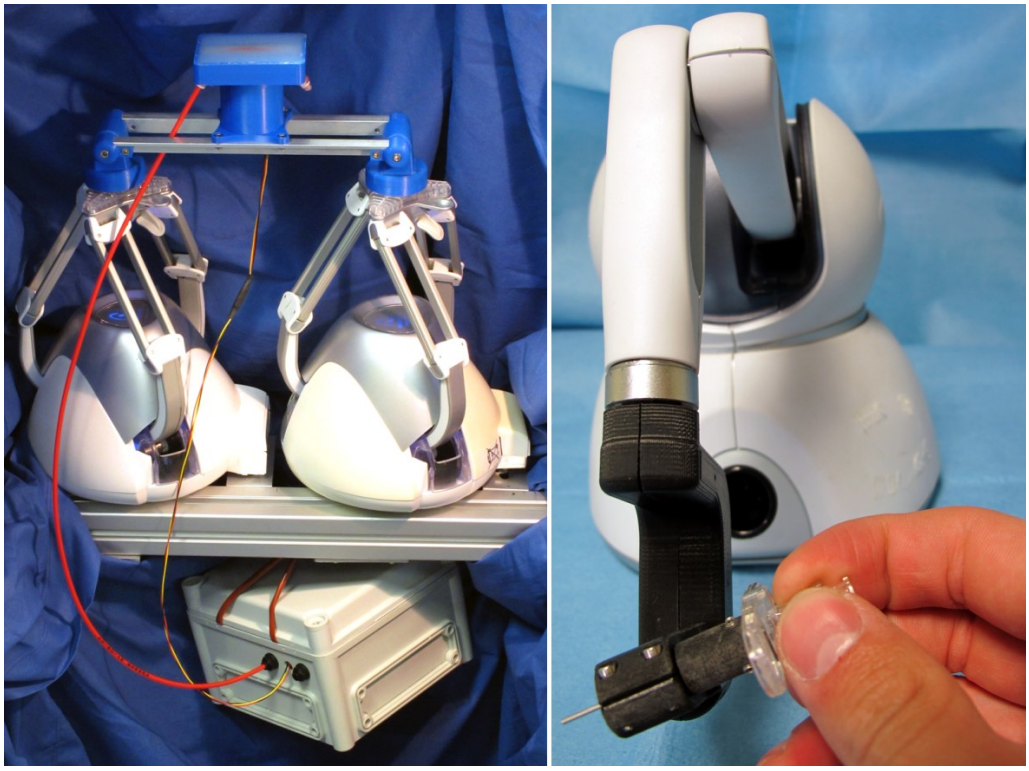


Figure 2.4: The tacttile and force feedback haptics interfaces used by PalpSim (Coles et al. 2011b): a visual-haptic simulator for femoral palpation and needle insertion.

One example is PalpSim (Coles et al. 2011b), an AR visio-haptic medical training system for femoral palpation and needle insertion. A custom tactile interface has

been fabricated for providing a pulse sensation in the artery of the virtual patient - see left image in Fig. 2.4. It consists of a silicon tray with a plastic tube embedded in it through which water is pumped with a pulsalate flow. The tray is mounted onto two modified NOVINT Falcons that enable the user to press down on the skin of the patient and feel an appropriate force response. A modified PHANTOM Omni (now called the GeoMagic Touch) force feedback device is also used to mount a real needle and impart accurate forces as the needle punctures the skin of the virtual patient into the artery - right image in Fig. 2.4.

Sutherland (Sutherland et al. 2013) developed an augmented reality haptic training simulator for spinal needle procedures using a PHANTOM Omni (now called the GeoMagic Touch) and patient specific computed tomography (CT) volume data. The novel aspect of this work is the simulation of Ultrasound images from CT volume using a deformed finite element model (FEM) to simulate the ultrasound-guided spinal needle procedures.

A particular problem when using a haptics device in AR is the superimposition of computer graphics (such as surgical instruments) to the haptics stylus to provide a realistic user experience. The PalpSim system used a chroma-key approach to mask out the physical haptics devices. Eck *et al* (Eck et al. 2014) proposed a calibration procedure for the precise co-location of visuo-haptics augmented reality by combining optical tracking with the information available on the angles of haptic stylus joints.

Saga and Deguchi (Saga and Deguchi 2012) reported the phenomenon that people tend to experience tactile perception if they receive force feedback in/against the direction of a moving/rotating 2D surface. Based on this phenomenon, Kim *et al* (Kim et al. 2014) proposed a robotic touch screen that could provide relative geometric information in the form of rotational force feedback which could be used in augmented reality applications. However, the speed and accuracy of this method needs to be improved to simulate realistic tactile feedback.

Spillmann (Spillmann et al. 2013) converted tactile feedback problem into an adaptive space warping problem by warping different organ geometries onto one

physical anatomical model to be used in a mixed-reality surgical training simulator. The virtual structures are deformed to adapt the physical model. Therefore users could feel the tactile feedback from the physical model while observing the mixed reality object. This method works to some extent, but modifying the tissue is only approximate, and the distortion sometimes could be unrealistic.

One of the latest technology approaches being explored for tactile devices is to use focused ultrasound. Hung *et al* (Hung et al. 2014) built an AR environment for locating a weak pulse or heart murmur sensation in a virtual patient. They used a hexagonal array of 271 ultrasound transducers focused on a particular location in space so that a hand can detect a tactile sensation at that location.

Haptic Augmented Reality (HAR) is also a term that has been used for augmenting haptics onto real objects. Jeon and Choi(Jeon and Choi 2010) proposed a framework that enables users to perceive augmented stiffness inside a real object supporting arbitrary exploration patterns such as tapping, stroking, and contour following. This technique was later extended into rendering a tumour in a silicone model(Jeon et al. 2012) to simulate the breast cancer palpation (Jeon et al. 2010a) (Jeon et al. 2010b).

2.3.1.3 Other Hand Held Controllers

BodyExplorerAR (Samosky et al. 2012) is a mannequin medical simulator based on projective AR. An IR pen is used to interact with the simulator to open, resize and move viewports providing windows into the body that can display dynamic anatomy.

Nataka *et al* (Nakata et al. 2012) use a smartphone as an AR remote control for a radiology review application. The phone is placed inside a hard case with a printed fiducial marker on its back for optical tracking. The phone's touch screen is then used for image manipulation with the AR environment being presented on a connected PC. The authors report greatly improved collaborative discussion and education.

2.3.1.4 Brain-Computer Interfaces

Recent advances in Brain-Computer Interfaces (BCI) are opening up new potential for use in medical AR. Lotte *et al* (Lotte et al. 2012) presented an overview of research activities to combine BCI and VR and concludes that BCI promises a more direct and intuitive way of interaction within a virtual environment. This is expected to also apply to AR. For example, "Superman-like X-ray vision" is a concept proposed by Blum *et al* (Blum et al. 2012b) for surgeons to see augmented pre-operative images onto the patient with display devices such as a head-mounted display (HMD). A BCI device would be used to switch on and off the augmented anatomy; and a gaze tracker would control the position of the focus window. However, this project is only in an early stage. Currently only electro-oculographic signals triggered by eye movements are used to control the display of the augmented view.

2.3.2 Display Applications

AR can be presented on a variety of display platforms with differing degrees of fidelity. The most popular display categories currently used in medical AR are summarised below.

2.3.2.1 Head Mounted Displays

There are currently two types of Head Mounted Displays (HMDs) for AR: video see-through and optical (Meng et al. 2015). A video see-through (video mix) HMD (Kato and Billinghurst 1999) captures video via a mono- or stereo-camera, and then overlays computer-generated content onto the real world video. The processed video is displayed on the screen of the HMD. However, this type of HMD isolates the user view from the real-world environment and is prone to high system latency due to the accumulation of video acquisition, image processing and rendering. The other type of HMD is the optical see-through (transparent) HMD, which allows users to directly see the real world while virtual objects are displayed through a transparent

Table 2.2: HMD-based medical AR applications Summary

Article	Purpose	HMD Type	HMD Device	Tracking
(Hanna et al. 2018)	Anatomic Pathology	Optical	Microsoft HoloLens	Manually Aligned
(Song et al. 2018)	Endodontics	Optical	Microsoft HoloLens	Marker
(Turini et al. 2018)	Hip Arthroplasty Training	Optical	Microsoft HoloLens	Marker
(Cosentino et al. 2017)	Radiotherapy	Optical	Microsoft HoloLens	Marker
(Kuhlemann et al. 2017)	Endovascular Interventions	Optical	Microsoft HoloLens	Magnetic Tracker
(Meng et al. 2015)	Veins Localization	Optical	Vuzix STAR 1200XL	Manually Aligned
(Chang et al. 2015)	Remote Surgical Assistance	Video	VUNIX iWear VR920	Optical Tracker + KLT*
(Hsieh and Lee 2015)	Head CT Visulization	Video	Vuzix Wrap 1200DXAR	KLT + ICP**
(Wang et al. 2015a)	Screw Placement Navigation	Optical	NVIS nVisor ST60	Optical Tracker
(Vigh et al. 2014)	Oral Implantology	Video	NVIS nVisor SX60	Optical Tracker
(Hu et al. 2013)	Surgery Guidance	Video	eMagin Z800 3D Visor	Marker
(Abe et al. 2013)	Percutaneous Vertebroplasty	Video	HMD by Epson	Marker
(Azimi et al. 2012)	Navigation in Neurosurgery	Optical	Goggles by Juxtopia	Marker
(Blum et al. 2012b)	Anatomy Visulization	Video	Not Mentioned	Gaze-tracker
(Tanaka 2010)	Cognitive Disorder Rehabilitation	Video	Canon GT270	No Tracking
(Wieczorek et al. 2010)	MIS Guidance	Video	Not Mentioned	Optical Marker
(Bretón-López et al. 2010)	Treatment of Cockroach Phobia	Video	HMD by 5DT	Marker
(Alamri et al. 2010)	Poststroke-Patient Rehabilitation	Video	VUNIX iWear VR920	Marker

* Kanade-Lucas-Tomasi algorithm (Tomasi and Kanade 1991).
** Iterative Closest Point algorithm (Best and McKay 1992).

Figure 2.5: Optical HMD to be used to navigate medical screws insertion (Wang et al. 2015a). Image courtesy of Huixiang Wang, Fang Wang, Anthony Peng Yew Leong, Lu Xu, Xiaojun Chen and Qiugen Wang

lens. Users will receive light from both the real world and the transparent lens and form a composite view of real and virtual object. The optical see-through HMD was first developed in the 1960s by Sutherland (Sutherland 1968). Half-silvered mirrors were used to enable users to see both the images from miniature CRTs and the real world simultaneously. With advances in display technologies, there appeared many other solutions for displaying images onto transparent lenses such as the reflective waveguide technique (Sarayeddine and Mirza 2013) (used by the Google Glass, and Epson Moverio^{TM} Series) and diffractive waveguide(Sarayeddine and Mirza 2013) (used by the Microsoft Hololens(Hempel 2015)). Another example is given in Figure 2.5 where the surgeon is wearing an optical HMD to assist with screw insertion (Wang et al. 2015a). Table. 2.2 provides a summary of the HMD-based medical AR applications that we have identified. Most of these medical applications use a video see-through HMD with marker-based tracking for simplification. The new generation of HMDs - Microsoft Hololens are becoming the mainstream for medical applications.

Rolland and Fuchs (Rolland and Fuchs 2000) comprehensively compared optical and video see-through approaches in the context of AR 3D medical visualization.

Their comparison focused on many aspects such as latency, resolution, distortion, field of view (FOV), viewpoint matching, cost, human-factors and perceptual issues. They conclude that hybrid solutions that use optical see-through technology for display and video technology for object tracking would play a key role in future developments.

Related to HMDs is the Virtual Retinal Display (VRD) (Furness III and Kollin 1995), a new technology that can scan low power laser light directly onto the human retina, creating bright, wide field-of-view, high contrast and high-resolution visual image (Stredney and Weghorst 1998). Together with the advantage of low power consumption, micro size, and no blockage of the visual field, VRD is well-suited for the next generation AR HMD solution such as the Magic Leap.

2.3.2.2 Mobile Displays

Other wearable AR devices are emerging that do not need to be attached to a PC and so improve the mobility of AR applications. Announced in 2012, Google Glass (Google 2018) was the first mobile head-mounted display designed in the shape of a pair of eyeglasses. Google stopped producing the Google Glass prototype in 2015 but this device demonstrated many potential uses in medical AR including pediatric surgery (Muensterer et al. 2014), tele-mentoring (Assad-Kottner et al. 2014), clinical education (Tully et al. 2015), forensic autopsy (Albrecht et al. 2013), remote electrocardiogram (ECG) interpretation (Jeroudi et al. 2015) and medical image retrieval (Widmer et al. 2014). However, the display screen of Google Glass is very small as it is designed for "providing" information not "superimposing" information onto the corresponding physical position. These applications can therefore only assist medical students, doctors or surgeons to a limited extent, such as taking first perspective photos or providing access to textual guidance information or medical images.

Microsoft HoloLens (Microsoft 2018) is a headset that consists of a pair of transparent combiner lenses that images can be projected onto, a depth camera to provide depth information, a video camera, and an inertial measurement unit (IMU) includ-

ing an accelerometer, gyroscope, and a magnetometer to provide speed, pose and direction information (Holmdahl 2015). With an integrated high-end CPU and GPU, HoloLens is actually a wearable computer running Windows 10, and is regarded as a revolutionary product in AR. Medical AR applications using the Hololens are expected to emerge quickly following the expected launch of this device in late 2016. Microsoft has demonstrated a possible medical application at the 2015 Microsoft Build Developer Conference (Holmdahl 2015).

Another product in this category about to be launched is the Meta 2 glasses (Meta Co. Redwood, CA). The Meta 2 will offer a high resolution screen with built in tracking of hand gestures. One of the early test uses of the Meta 2 glasses has been in the context of the Glass Brain project, a brain visualization implemented in Unity3D that displays source activity and connectivity, inferred in real-time from high-density EEG.

2.3.2.3 Spatial Augmented Reality

Spatial Augmented Reality (SAR) solutions make use of projectors, half-silvered mirrors, or screens to display augmented information directly onto a physical space without the need to carry or wear any additional display devices(Raskar et al. 1998). By augmenting information in an open space, SAR enables sharing and collaboration easier than with a single user HMD. There are currently three types of SAR solutions based on the display methods used: video see-through SAR, optical see-through SAR, and direct augmentation SAR (Carmigniani et al. 2011).

Video see-through SAR is a cost efficient screen-based solution that only requires a camera, computer and display screen. But has the disadvantage of high latency, bad image quality and the loss of 3D vision. Whereas for optical see-through SAR, these problems are tackled by using semi-transparent display to overlaying compute-generated objects onto real world. However, it is more expensive, need calibration and has low brightness/resolution for overlaid objects. Direct augmentation SAR do not use screen but directly projecting virtual information onto real-world objects. It is more natural, but has low accuracy.

In medical applications, Video see-through SAR is often used for minimally invasive surgery guidance where a ready-to-use video output is available to be augmented. Kang *et al* (Kang et al. 2014) proposed a video see-through SAR system for laparoscopic surgery. The live stereoscopic laparoscopic video was merged with real-time ultrasound images to create an ultrasound-augmented stereoscopic video stream, which could provide surgeons with the 3D spatial relationship between anatomical structures and the visual cues of important internal structures such as tumors, bile ducts, and blood vessels. Bernhardt *et al* (Bernhardt et al. 2016) presented an automatic endoscope localization algorithm for AR guidance in laparoscopic surgery. By estimating the position and direction of the endoscope tip in the volume data, the corresponding view from an intra-operative CT volume can be superimposed onto the endoscopic view to guide surgeons. Pico Lantern (Edgcumbe et al. 2015) is also aimed at laparoscopic surgery. It is a miniature projector developed for structured light surface reconstruction that can be inserted into a patients abdomen along with the laparoscopic camera probe. A known coded pattern is then projected onto the surface of organs to facilitate 3D surface reconstruction and further guidance information. The author reported the absolute error of 1.4 mm, 1.5 mm and 1.5 mm for plane, cylinder and kidney respectively by a mono laparoscope and a tracked Pico Lantern, which confirmed the accuracy of surface reconstruction.

Optical see-through SAR makes use of half-silvered (semi-transparent) mirrors, beam splitters or transparent screens to allow the user to simultaneously see the physical world and virtual objects in the same spatial position. Liao *et al* (Liao et al. 2010) proposed a 3D AR system for MRI-guided surgery based on a modified half-silvered mirror that could provide geometrically accurate 3D spatial images and reproduce motion parallax without using any supplementary eyeglasses or tracking devices - see Fig. 2.6. The auto-stereoscopic images were created by employing integral videography (IV) (Liao et al. 2004) technology, which could reproduce 3D images using a micro-convex lens array and a high-resolution high-pixel-density flat display. By this 3D optical see-through SAR system, surgeons can easily perceive

Figure 2.6: The Integral Videography stereo half-silvered mirror based SAR system for MRI-guided surgery. Image courtesy of Hongen Liao, Takashi Inomata, Ichiro Sakuma and Takeyoshi Dohi (Liao et al. 2010).

the depth of the IV image. Also based on a stereo half-silvered mirror system, Wang *et al* (Wang et al. 2014) presented an augmented reality navigation system for dental surgery using automatic marker-free image registration based on 3-D contour matching of teeth. Fritz *et al* (Fritz et al. 2012) evaluated the accuracy of a semi-transparent mirror based AR image overlay system in MRI-guided spinal injection procedures. The assessment results showed entry error of 1.6 $\pm$ 0.8 mm, angle error of 1.6° $\pm$ 1.0°, depth error of 0.7 $\pm$ 0.5 mm, and target error of 1.9 $\pm$ 0.9 mm, which could facilitate accurate MRI guidance for successful spinal procedures. Shi *et al* (Shi et al. 2012) remodeled a surgical microscope by inserting a beam splitter to allow users to see the microscope view with an augmented image from a pico-projector. Using this AR microscope, surgeons can observe virtual cues that track the movement of the tip of micro-surgical instrument, showing the desired position, and indicate the position error, which helps to maintain high performance and avoids the instrument drifting out of the workspace.

Direct augmentation SAR usually employs a projector or laser transmitter to

Figure 2.7: A projector based SAR anatomy learning system. Image courtesy of Adrian S. Johnson and Yu Sun (Johnson and Sun 2013).

project images directly onto the physical objects' surface. SARP (Spatial Augmented Reality on Person) (Johnson and Sun 2013) is a system that can project anatomical structures directly on user's dynamically moving body - see Fig. 2.7. Users can turn around and change pose to view the anatomy from different angles. A gesture-based interaction system was provided for users to select organs or switch between a muscular system or a skeletal system. Hochreiter *et al* (Hochreiter et al. 2015) used a P300 pico projector (AAXA Technologies, Tustin, USA) to project a head model onto a plastic human head shell. IR cameras were used to detect touch events on the head shell (see section 2.3.1.2). As the head shell is not a plain surface, the mapping relationships cannot be described in a parametric function. The author used a lookup table that contains correspondences amongst all coordinates to directly link the 3D graphics space, 2D projector space, 2D camera space and 3D touch space. ARCASS (Augmented Reality Computer Assisted Spine Surgery) (Wu et al. 2014) system is a projection-based AR system for spinal surgery assistance. The pre-operative 3D models constructed from CT images were superimposed onto patients with the help of markers, enabling the surgeon to see the patients' anatomy during spinal surgery. Experiments showed that the ARCASS system performed

well on phantoms and animal cadaver experiments, and also in clinical trials of orthopedic surgery.

SAR can also use a magic mirror technique where the user is presented with a mirror image of themselves on a screen. That image can then be augmented as desired. For example, Mirracle (Blum et al. 2012a) (Meng et al. 2013b) is an augmented reality magic mirror system for anatomy education. Ma *et al* (Ma et al. 2015) extended the magic mirror and enable a personalized anatomy model to be superimposed onto user's body in the mirror by detecting the user's height, body size, gender, and age. The evaluation study demonstrated that the average precision of the augmented reality overlay on the user body was 0.96 cm, also clinicians and students gave positive feedback towards the educational value.

Mirror-based AR systems have also been used in medical training (Stefan et al. 2014), education (Anderson et al. 2013) and rehabilitation (Erazo et al. 2014) applications. Finally, Mind-Mirror (Mercier-Ganady et al. 2014) is a virtual mirror that superimposes a virtual brain with a user's brain activity onto their own head. The brain activity is computed by EEG signals that are acquired in real-time and displayed using a mirror-based AR system.

2.3.3 Mobile AR

Traditional AR systems employ powerful computers to solve the heavy computation of camera pose estimation, target recognition, tracking and virtual object rendering; also, the bulky HMD is power-costly and must always be plugged into an electric socket. These factors greatly restrict the usage range of AR and limit its application. Recently, with the rapid development of micro-electronic technology, new mobile and wearable AR devices are becoming feasible (Höllerer and Feiner 2004), providing people with a more flexible and natural AR experience. They can also provide a low cost and portable solution which is expected to play a major role in medical/patient education and rehabilitation applications where accuracy of tracking is not critical.

2.3.3.1 Hand-Held Displays

Portable hand-held devices with integrated cameras (such as smartphones, tablets and PDAs) can provide a video see-through "window" that can overlay computer-generated graphics onto the captured real-time video (Wagner and Schmalstieg 2006) (Grandi et al. 2014). With the help of internal sensors such as accelerometers, gyroscopes, and magnetometers, three degrees-of-freedom of orientation can be acquired as shown in Fig. 2.8. Position information can be estimated by tracking markers or more advanced markerless tracking and Simultaneous Localization And Mapping (SLAM) techniques (see section 2.3.4).

Figure 2.8: Mobile AR for 3D visualization and interactive surgery planning. Image courtesy of Jeronimo G. Grandi, Anderson Maciel, Henrique G. Debarba and Dinamar J. Zanchet (Grandi et al. 2014).

Hand-held displays typically have small screen sizes, which can be a limitation with a restricted field-of-view and very limited depth perception, which may hide many details that are crucial in medical training and diagnosis. In addition, most mobile AR applications use marker-based tracking that rely on the position and focus of markers and will not work in poor lighting conditions. The display of patient specific data and 3D anatomical models will also be restricted on very small displays. Tablets are less prone to this disadvantage and can be used to create the

impression of a transparent display where the augmented real world is viewed as if looking through a window (Andersen et al. 2016). A depth camera and head tracker, however, should also be used to enable adjustments to be made to the rendered view through the transparent display so that it fits seamlessly with the real environment.

2.3.3.2 Smartphone and Tablet Applications

Table 2.3: Mobile Medical AR Applications

Article	Purpose	SDK	Device
(Andersen et al. 2016)	Surgical Telementoring	OpenCV	Project Tango
(Rantakari et al. 2015)	Personal Health Poster	Vuforia	Samsung Galaxy S5
(Kilgus et al. 2015)	Forensic Pathological Autospy	MITK*	Apple iPad 2
(Soeiro et al. 2015)	Brain Visulization	Metaio	Samsung Galaxy S4
(Juanes et al. 2014)	Human Anatomy Education	Vuforia	Apple iPad
(Kramers et al. 2014)	Neurosurgical Guidance	Vuforia	HTC Smartphone
(Noll et al. 2014)	Dermatology Education	Not Mentioned	Apple iPhone 4
(Garcia and Navarro 2014)	Ankle Sprain Rehabilitation	Vuforia	Apple iPad
(Virag et al. 2014)	Medical Image Visulization	JSARToolKit	Any device with browser
(Grandi et al. 2014)	Surgery Planning	Vuforia	Apple iPad 3
(Debarba et al. 2012)	Hepatectomy Planning	ARToolkit	Apple iPod Touch
(Choi 2011)	Stroke Rehabilitation	Not Mentioned	Android Smartphone

* Medical Imaging Interaction Toolkit

The popularity of smartphone and tablet devices are significantly increasing the accessibility of AR applications - many AR examples can be downloaded from both the App stores of Android and iOS platforms. In this section, three popular mobile AR Software Development Kits (SDKs) are introduced and medical AR applications are summarized in Table 2.3.

ARToolKit (DAQRI 2016) is a widely used open-source tracking library originally developed in 1999 and released by the University of Washington HIT Lab (Kato and Billinghurst 1999). After several iterations, ARToolKit has become one of the first AR SDKs for mobile devices running on Symbian, iOS and Andriod. The version 5.2 of ARToolKit has included some features that were previously only available in the professional licensed version. In particular, natural feature tracking was made availabl, which could be regarded as markerless tracking.

Vuforia (PTC 2018) is a commercial SDK for AR focused on mobile devices with over 175,000 registered developers. It has been used in apps for advertising, educa-

tion, games and tourism, etc., with more than 200 million app installed worldwide. Vuforia provides an API in C++, Java, Objective-C, and the .Net languages. It is also available as an extension to the Unity 3D games engine, which allows Vuforia-based applications to support iOS and Android. Like most of the AR SDKs, Vuforia uses computer vision technology to recognize and track planar images (marker and markerless images), but also supports text markers and simple 3D objects, such as boxes and cylinders in real-time. As a commercial SDK, Vuforia has some features that other AR SDKs do not have, such as Vuforia Cloud Recognition Service, which allows Vuforia-enabled applications to recognize image targets through a cloud database, giving developers the ability to update targets dynamically. The free version of the Vuforia software has a watermark on the camera view that cannot be removed and only 1,000 cloud-based recognitions can be performed.

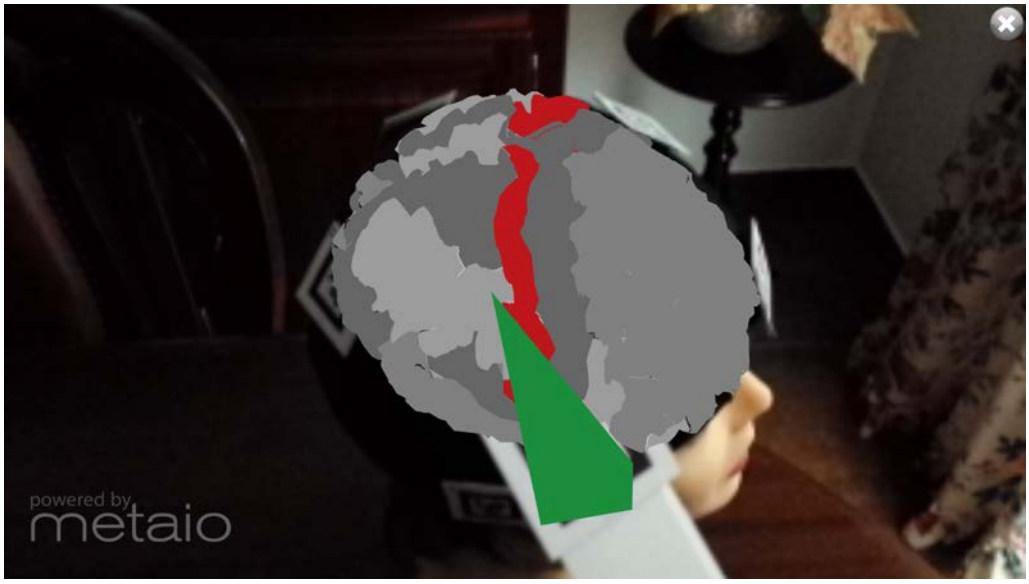


Figure 2.9: Brain Visulization on an AR Smartphone application using Metaio SDK. Image courtesy of José Soeiro, Ana Paula Cláudio, Maria Beatriz Carmo and Hugo Alexandre Ferreira (Soeiro et al. 2015).

Metaio (Apple 2018) produced a successful commercial SDK for AR applications. The Metaio SDK had a large community of developers with more than 1,000 customers and 150,000 users worldwide in 30 countries. Sale of the SDK ceased in May 2015 following the purchase of the company by Apple, Inc. (Cupertino, CA)

(Wakabayashi 2015). The Metaio SDK included a powerful 3D rendering engine in addition to plug-ins for Unity 3D. Metaio supported markerless 2D and 3D Tracking. Users could also create a complete AR scenario without specialized programming knowledge through a drag and drop interface called Metaio Creator. Fig. 2.9 is from one example of a medical AR application built using the Metaio SDK (Soeiro et al. 2015).

ARKit (Apple 2017) and ARCore (Google 2017) are the newly released mobile AR SDKs by Apple and Google. They are well calibrated with the sensors on their phones (e.g. camera and IMU) that provide best AR experience on mobile phones. The plane detections is also supported to delivery MR interactions.

As can be seen from Table 2.3, Vuforia is currently the most popular SDK for developing AR medical mobile applications. There is no particular bias towards either Android or iOS platforms. A recent study (Egui Zhu 2014) focused on the mobile AR in medical education has reported that AR with mobile technology could provide compelling, contextual, and situated learning experiences to medical students. Although some mobile AR applications are used in surgical planning and guidance, these are currently only prototypes built to demonstrate feasibility of using AR, and yet to gain regulatory approval.

2.3.4 Tracking

The tracking of real objects in a scene is an essential component of AR. This may involve detection and tracking of specific markers placed in the real world, which mark the location for the augmented content, or to use computer vision techniques to continuously track an object or person. The following summary is relevant to all applications of AR, not just medical.

2.3.4.1 Marker-based Tracking

Marker-based AR methods use distinctive specially designed markers that can be calibrated and tracked. During the early stages of AR development, marker-based

tracking was the most commonly used method to place a virtual object in real world. It is straightforward to implement and reliable for pose estimation. Several modalities of markers have been used, including optical trackers, magnetic trackers, and planar markers.

There are two types of optical tracking approaches: active and passive. An active marker is typically composed of infrared light-emitting diodes (ILED) that use near infrared light (approx. 850nm). Although this wavelength is invisible to human eyes, most camera sensors can detect it (Maletsky et al. 2007). Alternatively, passive markers are retro-reflective spheres that can be triangulated by identifying their position on a series of cameras mounted at different locations. Wiles *et al* (Wiles et al. 2004) compared the accuracy of active and passive optical tracking and reported that passive markers given slightly worse results than the active markers, but not significantly. Typical optical trackers usually can achieve an accuracy of 0.35mm (Wiles et al. 2004). Recently, smaller afocal optical markers (Chae et al. 2015) appeared to be extremely accurate (a position error of $219\mu m$ was reported) compared to conventional optical trackers. Optical trackers are commonly used in image-guided surgery (Ieiri et al. 2012) (Daly et al. 2010) (Zhang et al. 2013) (De Paolis and Aloisio 2010) (Wieczorek et al. 2010) due to the high accuracy that is needed. Some optical markers are specially made with iodine and gadolinium elements so that they can display high intensity in both X-ray/CT images and MR images (Maurer Jr et al. 1997). They can be attached to patient's skin or screwed into the bone of the skull to be visible both in CT/MR scanners and to an optical camera. This makes the registration between CT/MR images and video images easier in AR medical applications (Khan et al. 2006) (Nicolau et al. 2005). However, an optical tracking system also requires a free line-of-sight between the optical marker and the camera. Robotic methods have been attempted to overcome this restriction by using an optimization based control method to re-configure the optical tracker when the occlusion problem occurs (Wang et al. 2015b). Optical tracking is impossible, however, during laparoscopic surgery, when the camera is inside patient's body.

Magnetic tracking systems (Raab et al. 1979) offer an alternative approach without the direct view of tracking sensors. They use a magnetic-dipole source to generate a three-axis low-frequency quasi-static magnetic field. As sensors interact within the field, positional and orientation information relative to the source is generated. Kwartowitz *et al* (Kwartowitz et al. 2010) designed a phantom system to analyze the accuracy of two magnetic tracking systems used in image-guided surgery. The result shown that the average accuracy for all systems was greater than 3mm and gradually decays with distance from the field. The main working principles, error sources, tracking accuracy and robustness of the electromagnetic tracking technology in medicine were discussed in (Franz et al. 2014). Magnetic tracking in medical applications also lack robustness due to interference caused by diagnostic diagnostic devices or other ferromagnetic objects. An averaged accuracy of 1.0 mm, however, can be achieved by magnetic trackers in good environments (Franz et al. 2014).

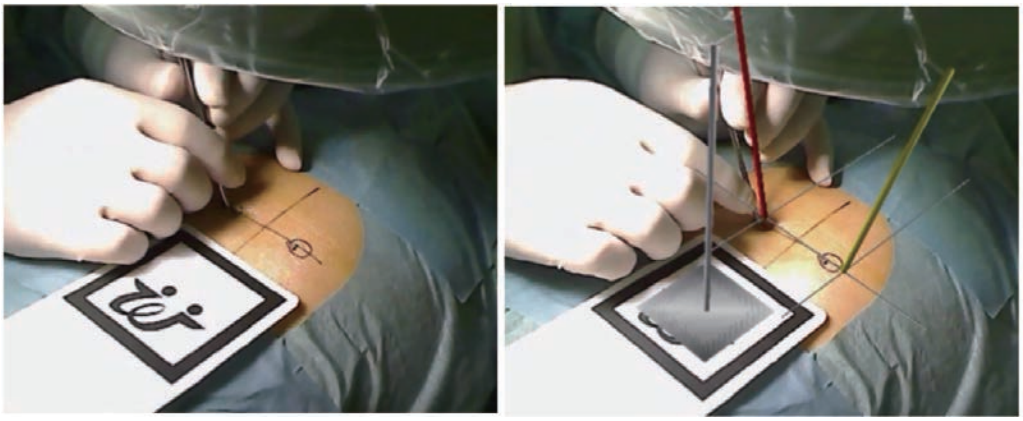


Figure 2.10: A marker-based AR 3D guidance system for percutaneous vertebroplasty; the augmented red line and yellow-green line indicate the ideal insertion point and needle trajectory. Image courtesy of Yuichiro Abe, Shigenobu Sato, Koji Kato, Takahiko Hyakumachi, Yasushi Yanagibashi, Manabu Ito and Kuniyoshi Abumi (Abe et al. 2013).

The use of planar markers, such as the example shown in Fig. 2.10, is one of the popular approaches for medical AR applications (Abe et al. 2013) (Loukas et al. 2013) (Lee et al. 2013) (HOSTETTLER et al. 2011) (Hu et al. 2013) (Azimi et al. 2012). Planar markers can be in many forms, such as concentric circles (Gatrell et al. 1992) (Calvet et al. 2012), square-shaped markers (Kato and Billinghurst

1999) (Fiala 2005) and barcode-based tags (Naimark and Foxlin 2002) (Rekimoto and Ayatsuka 2000). ARToolKit (DAQRI 2016) is a popular open source platform for rapid development of AR applications (see also section 2.3.3.2). ARToolKit markers consist of a black square border with a user defined image in the interior. The outside border is used to restore perspective distortion and estimate pose to match the interior pattern with templates. The usability, efficiency, accuracy, and reliability of marker based AR system, including ARToolKit, have been reported in (Zhang et al. 2002). The results showed that marker detection and decoding are fast and stable in ARToolKit, achieved the best scores in several aspects. Although being very useful for many applications, planar markers are prone to occlusion with a limited detection range and orientation. In addition, AR for minimally-invasive surgery is not possible to use these markers.

2.3.4.2 Markerless Tracking

In contrast to the marker based tracking system, markerless AR utilizes the real-world scenes and employs computer vision algorithms to extract image features as markers. The quality of markerless tracking, therefore, highly depends on lighting conditions, view angle and image distortion, as well as the robustness of the computer vision algorithm used. Markerless AR, however, enables a more natural AR experience with a wider range of applications, especially for medical applications such as the minimally-invasive surgery (Haouchine et al. 2015) (Elhawary and Popovic 2011).

The fundamental principles behind markerless AR are computer vision algorithms for feature detection and tracking, a process that usually contains three steps: detection of natural feature points; identification of discriminative descriptions (e.g. descriptors) of each feature point, as well as matching of feature points in image sequences. Feature points and descriptors are carefully selected and built to be invariant to affine transformation, light conditions, occlusion and noise to guarantee the robustness of the tracking performance in AR.

Scale Invariant Feature Transform (SIFT) (Lowe 2004) is one of the most widely

used and cited feature point detection and description algorithms. SIFT employs Histograms of Gradients (HoG) of neighbourhoods of feature points to construct 128-length descriptors as features. However, such a long descriptor does restrict its direct use in real-time AR applications. Later work (Ke and Sukthankar 2004) has shortened the feature matching time by applying Principle Component Analysis (PCA) to the descriptor and reduced the length of the descriptor to 36 for fast feature matching. However, features detected by this method are less distinctive than those by SIFT, and applying PCA can also slow down the computation. Similar to SIFT, Speed-up Robust Feature (SURF) algorithm (Bay et al. 2006) constructs a descriptor by summing Haar wavelet responses in the neighborhood of the feature points, resulting in faster performance than SIFT. Fig. 2.11 is taken from an example application that combines markerless SURF with an optical flow feature tracking (Plantefève et al. 2016). Further work using binary descriptors like Binary Robust Independent Elementary Features (BRIEF) (Calonder et al. 2010), Oriented FAST and Rotated BRIEF (ORB) (Rublee et al. 2011), Binary Robust Invariant Scalable Keypoints (BRISK) (Leutenegger et al. 2011) and Fast Retina Key- point (FREAK) (Alahi et al. 2012) can be much faster than HoG-based descriptors because comparing a binary string can be implemented by comparing the Hamming distance between them, which is equivalent to the sum of the XOR operation.

Mountney *et al* (Mountney et al. 2007) evaluated the performance of feature descriptors in computer vision to be used for tracking deformable soft tissue during MIS and concluded that best performing descriptors are Spin (Johnson 1999), SIFT, SURF, DIH (Ling and Jacobs 2005) and GLOH (Mikolajczyk and Schmid 2005). A novel probabilistic framework was proposed to combine multiply descriptors, which could reliably match significantly more features than by using individual descriptors. Experimental results showed that such fusion of descriptors could match a greater number of features even in the presence of large tissue deformation.

If images are acquired in a set of time steps, it is also possible to use optical flow (Horn and Schunck 1981) to compute the camera motion and track feature points (Mirota et al. 2011). Optical flow is defined as a distribution of apparent velocities

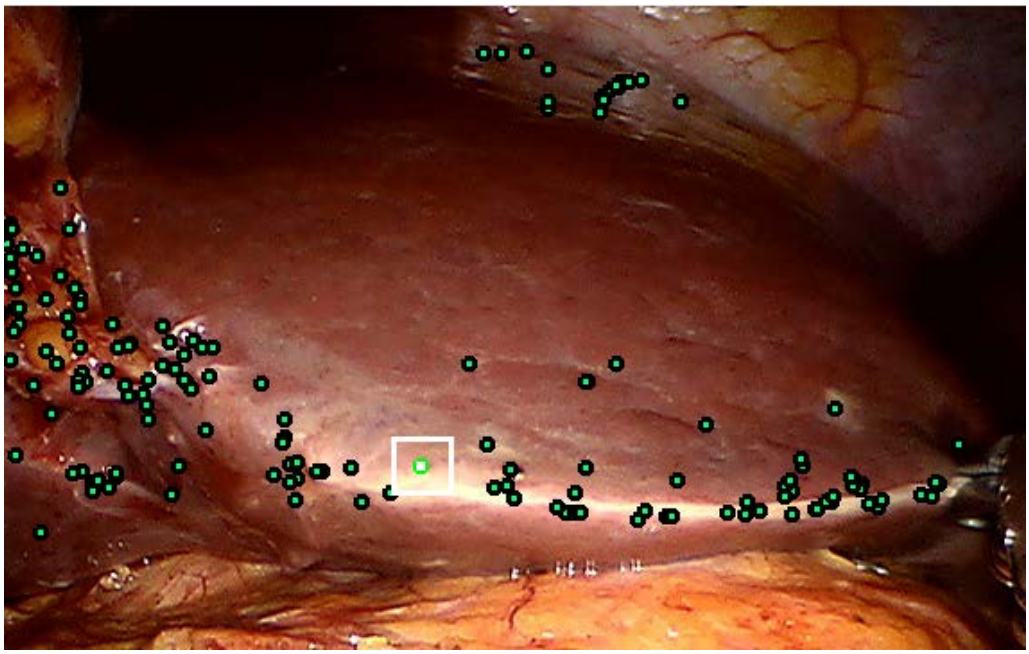


Figure 2.11: The markerless SURF feature on a liver rendered in the endoscopic video. Image courtesy of Rosalie Planteféve, Igor Peterlik, Nazim Haouchine, Stéphane Cotin (Plantefève et al. 2016).

of brightness patterns in an image (Horn and Schunck 1981), which can be used to track the movement of each pixel based on changes of brightness/light, which could be 10 times faster than SIFT feature construction (Lee and Hollerer 2008). Much of the literature in medical AR applications (Haouchine et al. 2015) (Haouchine et al. 2014) (Stoyanov et al. 2005b) (Stoyanov 2012) combines computationally expensive feature tracking with a light-weight optical flow tracking to overcome the performance issue of AR tracking. This approach only detects features periodically or only for initialization, while using optical flow to track feature points during the rest of the process (Lee and Hollerer 2008).

Particular challenges for medical AR are due to occlusions (from instruments, smoke, blood), organ deformations (respiration, heartbeat) (Puerto-Souza and Mariottini 2013) and the lack of texture (smooth regions and reflection of tissues). These factors mean that prior feature point detection and description methods designed for computer vision have limited capabilities in providing stable, consistent and real-time tracking for tissue surfaces(Yip et al. 2012). A custom framework for ac-

curately tracking tissues in surgery has therefore been investigated. in which, a STAR detector (derived from CenSurE (Agrawal et al. 2008)) was combined with a binary feature descriptor (BRIEF) to acquire robust salient features that can be tracked persistently on tissue surfaces in real-time. Evaluation results shown that the proposed framework outperforms other popular feature tracking algorithms with *In vitro* tissue experiments on kidney, heart and liver reported registration errors of only 1.3 to 3.3 mm. The hierarchical multi-affine (HMA) (Puerto-Souza and Mariottini 2013) feature-matching method was another approach specially designed for endoscopic images, which was proven (Puerto and Mariottini 2012) to be superior to the popular feature-matching methods frequently used in computer vision such as the nearest neighbor distance ratio (Lowe 2004) used in the original SIFT algorithm. HMA utilizes multiple affine transformations to pair features accurately between the two images. Tested in a large database with more than 100 pairs of MIS images obtained from real interventions, the HMA method outperformed the existing state-of-the-art methods in terms of speed, detection rate, and accuracy (Puerto-Souza and Mariottini 2013).

With the development of surgical instruments, stereo vision in MIS can be used to generate more robust and precise tracking. Chang *et al* (Chang et al. 2014a) presented a real-time visual odometry system for stereo endoscopy by dense quadrifocal tracking as shown in Fig. 2.12. By optimising the photometric energy function with respect to a camera pose constrained by the quadrifocal geometry, the movement of the endoscope can be estimated. The non-convex Tukey M-estimator is also used to rejects outliers for a robust tracking. Evaluation results were promising for synthetic, phantom and clinical data sequences.

2.3.5 Registration Techniques

Once the location for the augmented content has been determined, then this content (often computer-generated graphics) is overlayed or registered into the real word scene. Registration techniques usually involve an optimization step to minimize the difference (energy function) between virtual objects and real objects. For example,

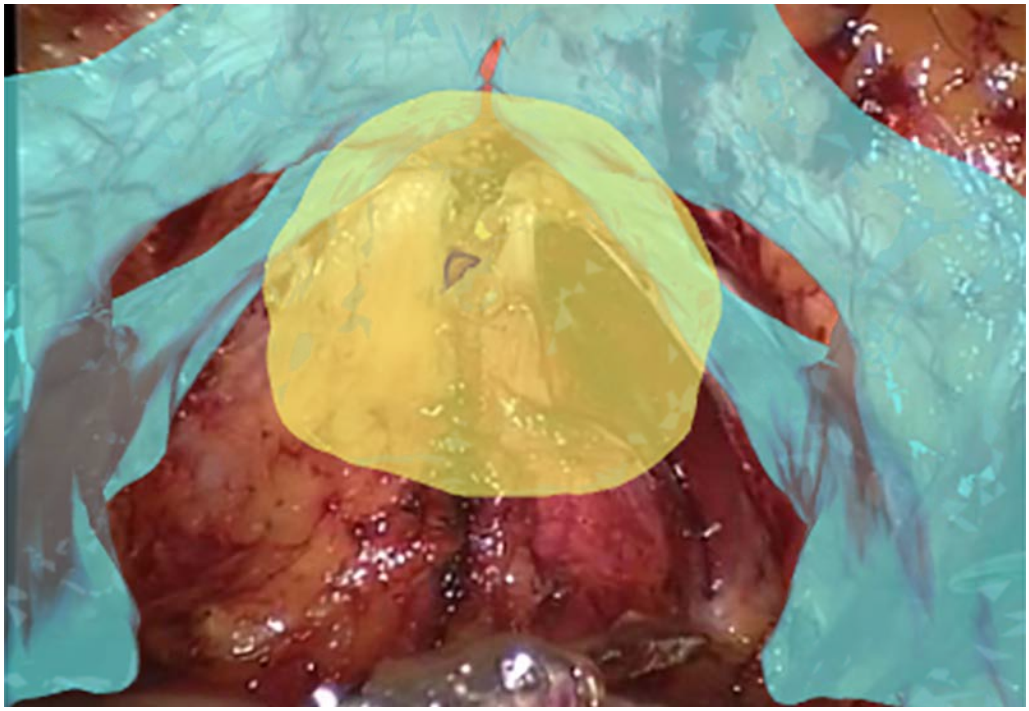


Figure 2.12: AR overlay onto laparoscopic images using a dense visual odometry method. Image courtesy of Ping-Lin Chang, Ankur Handa, Andrew J. Davison, Danail Stoyanov, Philip "Eddie" Edwards (Chang et al. 2014a).

using the 3D to 3D Iterative Closest Point (ICP) (Best and McKay 1992) technique, or other 2D to 3D algorithms (Markelj et al. 2012).

Wang *et al* (Wang et al. 2014) used an automatic marker-free image registration method in AR navigation of dental surgery. Patient image registration was achieved by matching a 3D teeth contour to a preoperative model derived from CT data. After the initial registration, the ICP algorithm was utilized to track the contours. The author reported that the overall mean error of the 3D image overlay was 0.71 mm, which was clinically satisfactory.

An improved ICP algorithm employed a weighting and perturbing strategy to increase robustness and noise resistance (Lee et al. 2012). The algorithm was tested to perform markerless registration between facial surfaces from preoperative CT images and stereo cameras, which allowed user to see the superimposed CT image on the corresponding position in the real word.

A biomechanical-based registration method was presented for pre- and intra-

operative 3D image fusion for laparoscopic surgery (Oktay et al. 2013). In this work, a gas insufflation model was built with pre-operative images using finite element model (FEM) to perform constrained registration with intra-operative images. Gradient ascent approach was then used to maximize the similarity measure by updating parameters of the FEM model. The validation on synthetic human CT and *in vivo* pig CT scans reported a mean registration error of 0.88mm to 1.75mm, which showed the applicability of the method in laparoscopic surgeries.

Registration of 2D to 3D (Markelj et al. 2012) (Chen et al. 2013) (Weese et al. 1997) is also effective for preoperative 3D data such as CT and MR images with intra-operative 2D data such as ultrasound (US), projective X-ray (fluoroscopy), CT-fluoroscopy, as well as optical images. These methods usually involve marker-based (external fiducial markers), feature-based (internal anatomical landmarks) or intensity-based methods that find a geometric transformation that brings the projection of a 3D image into the best possible spatial correspondence with the 2D images by optimizing a registration criterion (Markelj et al. 2012).

Registration of virtual anatomical structures within minimally-invasive surgery (MIS) video (Lamata et al. 2010) (Marescaux and Diana 2015) (Mirota et al. 2011) (Nicolau et al. 2011) (De Paolis and Aloisio 2010) is a much discussed topic. AR in MIS can significantly improve the quality outcomes of MIS. However, due to the problems of limited FOV, organ deformation, occlusion and no marker-based tracking possible, registration in MIS is still an unsolved problem.

A 3D to 3D ICP registration using image-based tracking to superimpose a 3D model onto laparoscopic images for partial kidney resection has been reported in (Su et al. 2009).The registration of 3D to 3D is performed assuming a rigid environment disregarding elastic deformations of kidney during surgery, which could be quite different from the pre-operative CT images.

Therefore, there are three big challenges when performing registrations in MIS, such as in laparoscopic liver surgery (Haouchine et al. 2015) (Plantefève et al. 2016):

1. Limited FOV of endoscope can only capture 30% – 40% of the whole liver surface.

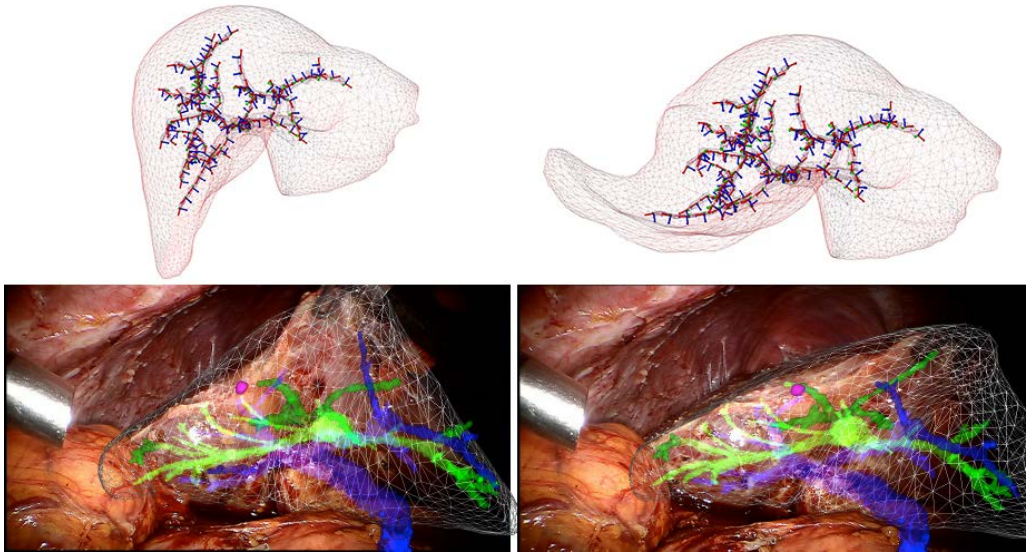


Figure 2.13: Registration of a physically-based liver model during minimally invasive liver surgery (Haouchine et al. 2013) (Haouchine et al. 2015). Top: Computer-generated heterogeneous liver model with the vascular network and the liver after deformation. Bottom: The superimposition of the real-time bio-mechanical model onto the human liver during surgery. Image courtesy of Nazim Haouchine, Jérémie Dequidt, Igor Peterlik, Erwan Kerrien, Marie-Odile Berger and Stéphane Cotin

2. Laparoscopic surgery requires inflation of abdomen with CO_2, which makes the shape of liver different from the pre-operative CT scan.

3. Liver is continuously moving due to both breathing and cardiac motions.

Referring to Fig. 2.13, a patient specific model was built from the pre-operative data and converted into a FEM model (with internal structures preserved such as tumours and vessels). A multi-step registration process was then used to align the FEM model with the real liver. The corresponding anatomical features were firstly manually labeled and ICP was used for the initial alignment. Temporal registration was then achieved by minimising an energy function between the internal force (generated by the displacement of points in FEM model) and the external force (the distance between the corresponding anatomical feature points). Feature points extracted by SURF and optical flow algorithms were used to track corresponding feature points for continuous elastic tracking. Although there were some failed registration cases at some positions, the *in vivo* test reported that the mean Hausdorff

distance between the FEM model mesh and the point cloud on the real liver was below 1.1 mm.

Simultaneous Localisation And Mapping (SLAM) (Dissanayake et al. 2001) was originally developed for the purpose of autonomous robot navigation in an unknown space. It has subsequently been applied to solve the problem of camera pose estimation in AR. In fact, AR has a very similar challenge as with robot navigation i.e. both need to get a map of the surrounding environment and locate current position and pose of cameras(Castle et al. 2008). However, applying SLAM on single handheld cameras (such as endoscopy cameras) is more complicated than with robot navigation as a robot is usually equipped with odometry tools and will move more steadily and slowly than a portable camera (Klein and Murray 2007). Non-linear SLAM was typically implemented as an Extended Kalman Filter (EKF) (Smith and Cheeseman 1986), where system noise was assumed Gaussian and non-linear models were linearized to suit the Kalman filter algorithm to solve the probabilistic SLAM problem (Bailey et al. 2006).

A monocular SLAM 3D model that combines an EKF with JCBB (Joint Compatibility Branch and Bound) (Neira and Tardós 2001) data association algorithm was proposed for endoscope image sequences (Grasa et al. 2009). A sparse abdominal cavity 3D map was created, and the motion of the endoscope was computed in real-time. This work was later improved (Grasa et al. 2011) by combining EKF monocular SLAM with 1-point RANSAC (Random Sample Consensus) (Civera et al. 2010) (Grasa et al. 2014) to deal with high outlier rate that occurs in real time and also to reduce computational complexity as shown in Fig. 2.14.

A Motion Compensated SLAM (MC-SLAM) framework for image guided MIS was presented (Mountney and Yang 2010a) , which predicted not only camera motions, but also employed an algorithm to learn a high-level model for compensating periodic organ motion, enabling estimation and compensation of tissue motion even when it is outside the camera's FOV. This framework was tested on both *ex vivo* and *in vivo* experiments; the augmented virtual tumor was consistently attached to the organ at the presence of both laparoscope and tissue motions. Based on

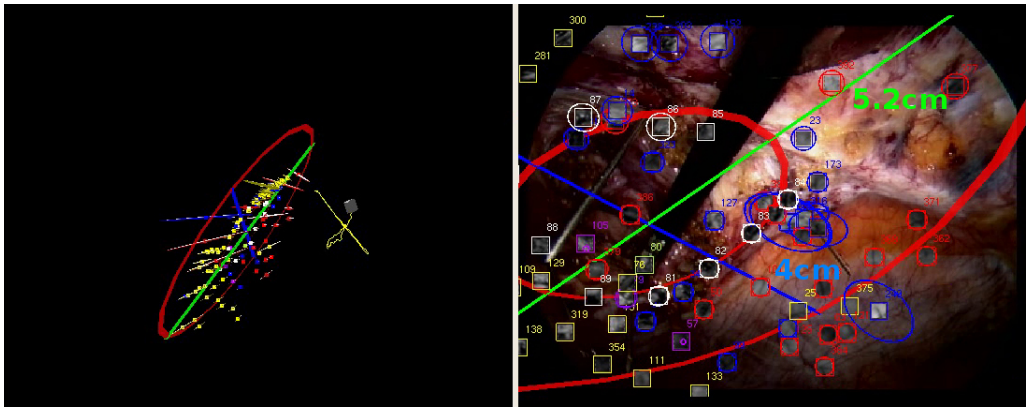


Figure 2.14: The monocular SLAM system used in MIS (Grasa et al. 2014). Left: Camera trajectory, 3D map and ellipses in 3D; Right: SLAM AR measurement, Map and ellipses over a sequence frame. Image courtesy of Óscar G. Grasa, Ernesto Bernal, Santiago Casado, Ismael Gil and J. M. M. Montiel.

this work, an AR framework for soft tissue surgery was proposed (Mountney et al. 2014). This approach utilized intra-operative cone beam CT and fluoroscopy as bridging modalities to register pre-operative CT images to stereo laparoscopic images through non-rigid biomechanically driven registration. In this way, manual alignment or fiducial marker were not required and also tissue deformation caused by insufflation and respiration during MIS was compensated while allowing AR overlays to be superimposed on laparoscope images.

2.4 Conclusion

In this Chapter, a data-mining based method is proposed for topic classification and trend analysis, along with a systematic technology review that covers recent AR technologies and applications. During the paper review process, I found the perception of the environment using camera is extremely important and useful for AR and can enable many advanced applications. Therefore, the research problem for this PhD project was identified as: vision-based dense surface reconstruction for geometry-aware AR. In next Chapter, the research problem and hypothesis will be presented following with a more focused literature review.

Chapter 3

Problem Statement & Literature Review

3.1 Problem Statement

In Minimally Invasive Surgery (MIS) such as laparoscopy procedures, an endoscope is inserted into the patients' body to reach internal organs through very small incisions. By performing MIS, patients can benefit from small incisions, less pain, low risk of infection and quick recovery time. However, while MIS offers considerable advantages to the patient, they are also imposing several additional difficulties on the surgeons. Different from open surgery, where the organs are exposed to the surgeon, in MIS, complex operations were carried out through the 2D visual display of video streams from the endoscopic camera. The limited FOV, the lack of depth perception and unnatural interaction will limit the performances of surgeons.

There were 8061 device malfunctions, 1391 injuries and 144 deaths recorded among a total of 1.7 million MIS procedures carried out between 2000 to 2013 (Alemzadeh et al. 2016). The numbers are still rising as MIS procedures become common. During MIS, surgeons have to find the target and perform complex operations under a small FOV endoscopic video stream. Errors are mainly due to disorientation, hand-eye disalignment and the difficulty of identifying surgical sites through mentally matching the laparoscopic view with pre-operative images (Kim

et al. 2012). Therefore, automatically overlaying surgical target locations on laparoscopic images is a highly important research topic in order to assist surgeons during MIS. Not only the solution will benefit surgeons, but also for patients due to reduce the likelihood of surgical errors and complications.

Recent advances in computer hardware and software technologies have enabled the use of computer vision techniques for MIS scene guidance and information augmentation, for example, using AR guidance systems to visualized pre-operative CT images (Kim et al. 2012) (Su et al. 2009), AR-based tumour visualization in laparoscopic surgery (Bourdel et al. 2017), and AR mapping for anatomy structures in liver MIS surgery (Haouchine et al. 2013) (Haouchine et al. 2015). In MIS, however, the luminance changes dramatically and also the movement of endoscope can also change rapidly during the insertion and the extrusion, which impose some particular challenges. The traditional tracking method for AR used in MIS scene usually involves the method of feature points tracking, such as the Scale-Invariant Feature Transform (SIFT) (Kim et al. 2012), Speeded Up Robust Features (SURF) (Kumar et al. 2014), Optical Flow tracking (Planteféve et al. 2016) and other approaches specifically designed to work with soft tissues to account for scale, rotation and brightness (Mountney and Yang 2008). However, these invariant descriptors are aimed at 2D tracking, and the depth perception issue remains unsolved (i.e. information regarding the depth of elements within a scene has not been recovered). In these algorithms, the selected feature points extracted from vision algorithms must be within the field of view. With these drawbacks, traditional feature tracking methods can severely affect the precision of virtual guidance, especially in the surgical scenes where accuracy is paramount.

A stereo endoscope will improve the depth perception problem, and such systems are now integrated into robotic systems (e.g. the da Vinci system from Intuitive Surgical, Inc.) or with the use of proprietary stereo cameras. 3D depth information can be recovered using the disparity map from rectified stereo images during a laparoscopic surgery (Stoyanov et al. 2004) (Stoyanov et al. 2005a), so that a 3D reconstruction using dense point clouds captured in the laparoscopic scene can

be achieved by a propagation method (Stoyanov et al. 2010) and/or a cost-volume algorithm (Chang et al. 2013). Stereo vision based reconstructions, however, can only recover the structure within localised frames without a global overview of the entire scene, which is very sensitive to noise and luminance changes (as mentioned above). When using robotic surgical systems, surgeons need to wear 3D glasses or view through a binocular interface. In addition, compared with monocular endoscopes, stereo endoscopic surgery is still too expensive to be widely used in practice. Therefore, there is an urgent need for the technology of on-the-fly building global surface using monocular camera (e.g. during monocular MIS or games with monocular camera). Providing this global structural information will not only improve the perception of depth in monocular surgical procedures, but also have promising applications when combined with AR technology for providing geometry-aware AR guidance.

3.2 Research Hypothesis

Here, the research hypothesis of this thesis is given:

It is hypothesized that, *the dense 3D surface of objects such as internal organs can be reconstructed from a live video feed using a novel vision-based approach, enabling more Augmented Reality applications in MIS and other fields, and so improving the user experience.*

3.3 Camera Tracking for AR

3.3.1 Feature-based 2D Tracking

Recent advances in computer hardware and software technologies have facilitated the use of computer vision techniques for MIS scene guidance and information augmentation. For example, AR guidance systems have been used to visualize preoperative CT images (Kim et al. 2012) (Su et al. 2009), for tumour AR visualization in laparoscopic surgery (Bourdel et al. 2017) and anatomy structures AR mapping

in liver MIS surgery (Haouchine et al. 2013) (Haouchine et al. 2015). There are, however, some particular challenges faced with AR in MIS. The luminance changes dramatically and an endoscope can move rapidly during insertion and extrusion.

The traditional MIS AR approaches usually employ feature points tracking methods for information overlay. Feature based 2D tracking methods such as Kanade-Lucas Tomasi (KLT) features (Du et al. 2015) (Planteféve et al. 2016), Scale-Invariant Feature Transform (SIFT) (Kim et al. 2012), Speeded Up Robust Features (SURF) (Kumar et al. 2014), Optical Flow tracking (Planteféve et al. 2016) or even those methods specifically designed to cater for the scale, rotation and brightness of soft tissue (Mountney and Yang 2008) have several major drawbacks for AR. As these invariant descriptors are designed for 2D tracking, the information regarding the depth within a scene has not been recovered and the selected feature points extracted from vision algorithms must be within the field of view. This result in the lack of global information in AR augmentations and the difficulty of acquiring these features pre-operatively. Also, traditional feature tracking methods can severely affect the precision of procedure guidance, especially in surgical scenes where the accuracy is paramount.

3.3.2 SLAM-based 3D Tracking

Recently, the maturity of the method of simultaneous localization and mapping (SLAM) designed for robot navigation in unknown 3D environments has opened up new opportunities for developing novel endoscopic camera tracking approaches in MIS. SLAM-enabled systems make it possible to estimate the 3D structure of the MIS scene from a moving endoscope camera and simultaneously track the pose of the camera. The scenario of the camera tracking and scene reconstruction in endoscopic surgeries is similar to that of a typical SLAM application in robotic vision, albeit with additional challenges. SLAM-enabled AR systems not only improve the usability of AR in MIS due to no optical or magnetic tracking devices to obstruct the surgeons' view, but they also offer greater accuracy and robustness compared with traditional feature-based AR systems. Based on the tracking methods, there

are two types of SLAM systems: Direct-based SLAM and Feature-based SLAM.

Direct-based SLAM algorithms compare pixels (Engel et al. 2014) or reconstructed models (Chang et al. 2014b) (Turan et al. 2017) of two images to estimate camera poses and reconstruct a dense 3D map by minimising the photometric errors. However, direct methods are more likely to fail when dealing with deformable scenes or when the illumination of the scene is inconsistent. Feature-based SLAM systems (Klein and Murray 2007) (Mur-Artal et al. 2015) only compare a set of sparse feature points that are extracted from images. These methods estimate camera poses by minimising the re-projection error of the feature points. Therefore, feature-based SLAM methods are more suitable for MIS scenes due to it's tolerance to illumination changes and small deformations.

Feature-based SLAM such as EKF-SLAM has been used with laparoscopic image sequences (Mountney et al. 2006) (Mountney and Yang 2009) (Grasa et al. 2014) and a further motion compensation model (Mountney and Yang 2010a) and stereo semi-dense reconstruction method (Totz et al. 2011) were integrated into the EKF-SLAM framework to deal with periodic deformation. However, the accuracy of EKF-SLAM tracking is not guaranteed and prone to inconsistent estimation and drifting due to the linearization of motion model and sensor models approximated by a first-order Taylor series expansion. The first keyframe-based SLAM – PTAM (Parallel Tracking and Mapping) (Klein and Murray 2007) was a breakthrough in visual SLAM and has been used in MIS for stereoscope tracking (Lin et al. 2013). The extension of PTAM – ORBSLAM (Mur-Artal et al. 2015) has also been tested on endoscope videos with map point densifying modifications (Mahmoud et al. 2016), but the loss of accuracy still exists. Furthermore, since feature-based SLAM systems can only reconstruct maps based on sparse landmark-points that barely describe the detailed 3D structure of the environment, the augmented AR content has to be mapped onto a plan through planar detection algorithms such as Random Sample Consensus (RANSAC) (Lin et al. 2013). Although feature-based SLAM is computationally efficient, different to real-life environments, in MIS scenes, flat surfaces are rare and organs and tissues do have smooth and curved surfaces, hence, resulting in

inaccurate AR content registration. One example is the inaccurate labelling and measurement of tumour size without accurate surface fit for information overlay, which can be dangerous and misleading during MIS.

3.4 3D Dense Surface Reconstruction

Another problem for feature-based based SLAM method is the lack of dense map. The saved feature points are usually very sparse for fast computation, which makes the map unexplainable and useless for further surface based applications. With external devices apart from camera, there are several practical solutions for 3D structure reconstructions, such as Constraint-based factorization methods (CBFM) (Wu et al. 2007), but external tracking devices are needed to provide the surgical instruments position. Lin *et al* (Lin et al. 2017) combined structured lighting with structure from motion for monocular endoscopic image reconstruction. Although special optical probe is needed, better reconstruction density and robustness are achieved with the extra benefit of super spectral resolution. The use of external device makes the reconstruction more accurate and reliable, but can also reduce its usability in practice. With the development of computer vision algorithms, there are several approaches to achieve purely image-based reconstruction.

3.4.1 Stereo Depth Estimation

The problem of stereo images depth estimation has been well studied for a long time (Barnard and Fischler 1982) (Scharstein et al. 2001). With the theory of epipolar constraint, accessing depth from stereo images can be regarded as a well-posed problem when ignoring the occlusions and depth discontinuities. Many stereo vision algorithms managed to achieve comparable results to ground truth depth acquired from depth sensors (Hirschmuller 2008) (Kendall et al. 2017). In laparoscopic surgery, 3D depth information can then be recovered using the disparity map from rectified stereo images (Stoyanov et al. 2004) (Stoyanov et al. 2005a) (Chen et al. 2017c), so that a 3D reconstruction using a dense point cloud of the laparoscopic

scene can be achieved by a propagation method (Stoyanov et al. 2010) and/or a cost-volume algorithm (Chang et al. 2013).

3.4.2 Monocular Depth Estimation

In contrast, estimating depth from monocular images is an ill-posed problem that is inherently ambiguous (Eigen et al. 2014), and many research efforts have been devoted to the problem of monocular image depth estimation. One of the classic methods is Shape from Shading (SFS) (Zhang et al. 1999), which is based on the gradual variation of shading as a cue to estimate the shape and depth. The SFS is also used for inferring the depth from monocular endoscopic images (Visentini-Scarzanella et al. 2012). However, SFS has a strict prior assumption of Lambertian reflectance, uniform color and texture, and fixed light source direction, which are not applicable to most of the images in the real world. Especially in MIS environment, the diffuse and specular reflection do exist due to the complex surface conditions of different tissues which will severely affect the accuracy of shape from shading. Saxena et al (Saxena et al. 2006)(Saxena et al. 2007)(Saxena et al. 2008)(Saxena et al. 2009) used Markov Random Field (MRF) incorporated with multiscale image features to learn monocular cues in a supervised manner. However, the hand-craft local features used in these approaches limit the expressive power of supervised learning, and lack a global contextual understanding of the scene for learning consistent depth.

3.4.3 DCNNs based Monocular Depth Learning

More recently, DCNNs (Eigen et al. 2014) (Eigen and Fergus 2015) are introduced to solve the challenge of monocular depth estimation problem, and has pushed the state-of-the-art forward in this area. Building on the success of this approach, several improvements have been made by incorporating probabilistic models such as Conditional Random Fields (CRFs)(Li et al. 2015) (Liu et al. 2014) (Hua and Tian 2016) (Liu et al. 2016) (Xu et al. 2017), advanced network structures such as Resnet

(Laina et al. 2016), two-streamed networks (Li et al. 2017b), multi-task joint training (Ladický et al. 2014) (Eigen and Fergus 2015) (Wang et al. 2015c) (Mousavian et al. 2016) (Yan et al. 2018) and novel loss functions such as sparse supervision (Kuznietsov et al. 2017), relative depth (Zoran et al. 2015)(Chen et al. 2016) and depth as classification (Cao et al. 2017). Impressive as these works are, ground-truth depth data are still needed for the supervision of training these DCNNs. However, it is still very difficult to build a large dataset with groundtruth for surgical scenes.

3.4.4 Unsupervised Monocular Depth Learning

Driven by DCNNs, view synthesis technology (Fitzgibbon et al. 2003) has proven to be effective on synthesizing new views by sampling pixels from existing views (Zhou et al. 2016) (Flynn et al. 2016), which enables novel frameworks of unsupervised learning of monocular depth from stereo pairs, e.g., Deep3D (Xie et al. 2016), Garg *et al* (Garg et al. 2016). The works by Godard *et al* (Godard et al. 2017) and Zhou *et al* (Zhou et al. 2017) advanced the networks by incorporating left-right consistency and pose estimations. However, a common weakness of these approaches is the use of pixel-wised photometric loss (L1-norm) to construct loss functions to guide the back-propagation process. Gradients are derived from the pixel intensity difference (Zhou et al. 2017), which will lead to ambiguous gradients in texture-less areas and also in the regions that contain the mixture of thin structures and texture-less areas. Although multi-scale and smoothness loss functions are used to prevent such issue (Garg et al. 2016) (Godard et al. 2017) (Zhou et al. 2017), the result is still not desirable and gradients are still likely to converge to local minimums due to the ambiguous pixel-wise loss.

3.5 Summary

In summary, the camera tracking and surface reconstruction technologies are both well developed in recent years. Camera tracking is the main enabling technology for AR, and there are many existing methods that are suitable for persistent and large-

scale camera tracking for AR. Whereas surface reconstruction technology is seldom applied in the scope of AR, and there lack a systematic framework for on-the-fly incrementally building global surfaces during AR. This technology is essential for geometry-aware AR and the more advanced context-level AR.

Chapter 4

Monocular-based Online Dense Surface Reconstruction for GA-AR

4.1 Introduction

In Minimally Invasive Surgery (MIS), medical procedures are technically demanding, and the difficulty is exacerbated by well-known issues and restrictions associated with MIS, such as the limited field of view (FOV), lack of hand-eye alignment and orientation, and the lack of stereoscopic depth perception in monocular endoscopy. Augmented Reality (AR) technology can help overcome these limitations by overlaying additional information onto the real scene such as annotations at target surgical locations (Kim et al. 2012), labels (Su et al. 2009), measurements of tumour sites (Bourdel et al. 2017) or even overlay a 3D reconstruction of anatomy (Haouchine et al. 2013) (Haouchine et al. 2015).

Despite recent advances in powerful miniaturized AR hardware devices and improvements on vision based software algorithms, many issues in medical AR remain unsolved. In particular, the dramatic changes in tissue surface illumination and tissue deformation as well as the rapid movements of the endoscope during insertion

and extrusion, all give rise to a set of unique challenges that call for innovative approaches. As with any other technological assisted medical procedure, the accuracy of AR in MIS is paramount.

The miniaturized devices in MIS mean that the Field of View (FOV) captured by a monocular endoscopic camera is usually very small, for example, only 30% to 40% of the whole liver surface is visible in one frame at one time (Planteféve et al. 2016). Traditional AR approaches (i.e. marker-less AR) for MIS are mainly based on feature tracking methods that require those selected feature points to be within the field of view (Haouchine et al. 2013). Given the restricted FOV, the algorithmic limitations of traditional methods can severely affect the precision of AR for procedure guidance. Our proposed geometry-aware AR framework addresses the issue by providing global 3D geometric information of the entire surgical scene so that the information overlay does not depend on the frame by frame local feature extractions, hence, greatly improving the reliability of AR augmentations.

Studies have shown that a typical human uses 14 visual cues to perceive depth, and 11 of the 14 cues do not require binocular vision (Dunkin and Flowers 2015). For example, depth information can be inferred in monocular vision through occlusions, motion parallax, shadows and texture gradient, and relative size and familiar size etc. The cognitive process of monocular vision enables surgeons to perform laparoscopic under a 2D environment (Mistry et al. 2013). However, monocular depth cues can only roughly estimate the general depth between objects, the accurate distance between objects cannot be perceived (Saunders and Backus 2006). Although examples of stereoscopic endoscopes do exist, they are not commonly accessible in medical practice (Velayutham et al. 2016) (Wagner et al. 2012). We address the aforementioned challenges by providing accurate geometric measurements and artificially generating depth cues through AR technology, which are important improvements in monocular endoscope environment for surgeons to carry out complex procedures. In our AR framework, the distance between objects can be deciphered by relative sizes of AR labels and annotations.

A stereo endoscope can provide stereoscopic vision and such systems are currently

available and often integrated into robotic systems (e.g. the da Vinci system from Intuitive Surgical, Inc.). 3D depth information can then be recovered using the disparity map from rectified stereo images during a laparoscopic surgery (Stoyanov et al. 2004) (Stoyanov et al. 2005a) (Chen et al. 2017c), so that a 3D reconstruction using a dense point cloud of the laparoscopic scene can be achieved by a propagation method (Stoyanov et al. 2010) and/or a cost-volume algorithm (Chang et al. 2013). Stereo vision based reconstructions, however, can only recover the structure of a local frame without a global overview of the scene, and are very sensitive to noise and luminance changes. Surgeons have to wear 3D glasses or use a binocular viewer on the robotic surgical system. In addition, stereo endoscopic surgery is still too expensive and yet to be widely used in practice. Hence, providing depth cues in monocular endoscope operations will have a significant impact on the accuracy of surgical procedures.

While Minimally Invasive Surgery (MIS) offers considerable benefits to patients, it also imposes big challenges on a surgeon's performance due to well-known issues and restrictions associated with the field of view (FOV), hand-eye misalignment and disorientation, as well as the lack of stereoscopic depth perception in monocular endoscopy. Augmented Reality (AR) technology can help to overcome these limitations by augmenting the real scene with annotations, labels, tumour measurements or even a 3D reconstruction of anatomy structures at the target surgical locations. However, previous research attempts of using AR technology in monocular MIS surgical scenes have been mainly focused on the information overlay without addressing correct spatial calibrations, which could lead to incorrect localization of annotations and labels, and inaccurate depth cues and tumour measurements. In this chapter, we present a novel intra-operative dense surface reconstruction framework that is capable of providing geometry information from only monocular MIS videos for geometry-aware AR applications such as site measurements and depth cues. We address a number of compelling issues in augmenting a scene for a monocular MIS environment, such as drifting and inaccurate planar mapping.

In this chapter, we present a novel method and a computational framework to

achieve accurate geometry-aware AR through: (i) extracting 3D depth information from camera motions and 3D surface reconstructions; and (ii) using AR technology to fuse rich 3D structural information with a monocular endoscope video stream, such that accurate spatial information in the scene can be derived from the scene geometry, and artificial depth cues can be provided based on the collaboration of the 3D spatial scene with the real-time video streams (i.e real-virtual overlay and simultaneous mapping). To this end, we explore the potential of the state-of-the-art SLAM framework by modifying and fine-tuning the algorithm for endoscopic camera tracking and mapping, so that the balance between point cloud density and computational performance can be achieved. A 3D surface reconstruction method based on the Moving Least Squares (MLS) smoothing and the Poisson surface reconstruction algorithms are proposed to recover a smooth surface from the unstructured sparse map points extracted from the MIS scene. Simulated laparoscopic sequences generated in a 3D modelling package have been used to evaluate the performance of the proposed framework in terms of robustness of the camera tracking and the accuracy of the surface mesh reconstruction.

The obtained global geometric information can be seamlessly integrated into our proposed AR framework, which is capable of achieving AR augmentations at the correct depth and detailed accurate surface measurements. Our method provides new possibilities for novel geometrically informed AR augmentations in monocular endoscopic MIS, including accurate annotations, labels, tumour measurement and artificial depth cues at correct depth locations that are demonstrated with two example applications: i.e. generations of artificial depth cues and the surface measurements of target sites in MIS.

We demonstrate the clinical relevance of our proposed system through two examples: a) measurement of the surface; b) depth cues in monocular endoscopy. The performance and accuracy evaluations of the proposed framework consist of two steps. First, we have created a computer-generated endoscopy simulation video to quantify the accuracy of the camera tracking by comparing the results of the video camera tracking with the recorded ground-truth camera trajectories. The accuracy

of the surface reconstruction is assessed by evaluating the Root Mean Square Distance (RMSD) of surface vertices of the reconstructed mesh with that of the ground truth 3D models. An error of 1.24mm for the camera trajectories has been obtained and the RMSD for surface reconstruction is 2.54mm, which compare favourably with previous approaches. Second, *in vivo* laparoscopic videos are used to examine the quality of accurate AR object placement, and the creation of depth cues. These results show the potential promise of our geometry-aware AR technology to be used in MIS surgical scenes.

The results show that the new framework is robust and accurate in dealing with challenging situations such as the rapid endoscopy camera movements in monocular MIS scenes. Both camera tracking and surface reconstruction based on a sparse point cloud are effective and operated in real-time. This demonstrates the potential of our algorithm for accurate AR localization and depth augmentation with geometric cues and correct surface measurements in MIS with monocular endoscopes.

4.2 Methodology

The flowchart in Figure 4.1 demonstrates our intra-operative MIS AR framework. As can be seen from Figure 4.1 (a), the endoscope is inserted into the patient abdominal cavity, which is inflated with carbon dioxide gas to create the pneumoperitoneum. Image sequences captured by the moving endoscopic camera are the input to our AR framework as shown in Figure 4.1 (b). The SLAM algorithm recovers the camera pose and generates an unorganized sparse point cloud. 3D geometric information is then built based on the point cloud by our proposed surface reconstruction framework. The dense surface mesh is then aligned with the input image sequences via a camera space transformation. Finally, the virtual object can be displayed onto the reconstructed surface to provide both depth cues and any virtual augmentation at the correct depth.

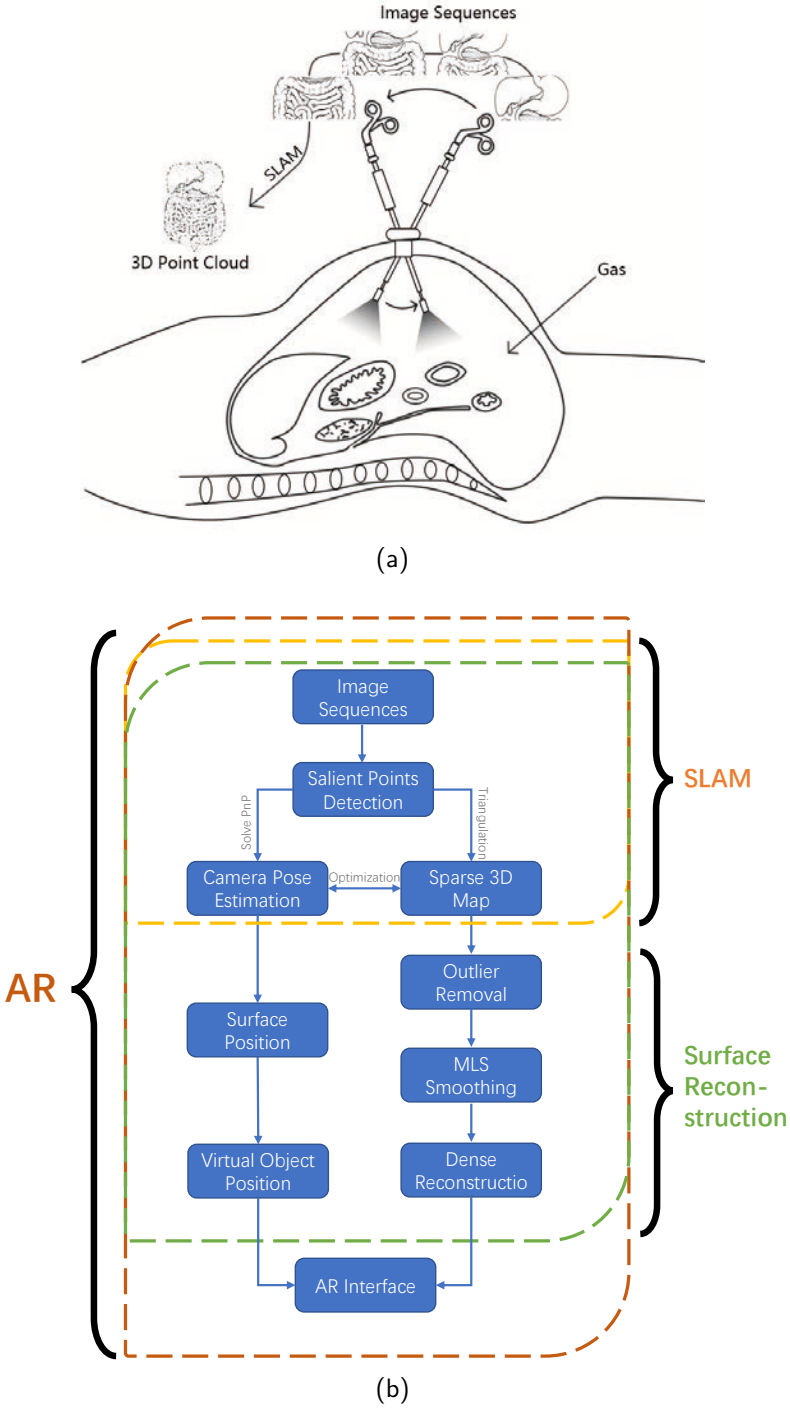


Figure 4.1: (a). A moving monocular endoscopic camera can capture a series of image sequences which can be used to build a 3D sparse point cloud by using a SLAM system. (b) The flowchart of our proposed AR framework.

4.2.1 Introducing of the Surface Coordinate

The difference between our approach and the feature based tracking method used in the previous AR work in MIS (Kim et al. 2012) (Haouchine et al. 2015) is shown in Figure 4.2, illustrating the use of different coordinate systems. The endoscope is represented as a probe in the camera coordinate system and p_c is a 2D point in the camera's view and the virtual object (the face) to be displayed. Figure 4.2 (a) shows the feature tracking based AR environment. When the feature is detected and tracked, the virtual object will be placed in the feature coordinate system. Assuming P_f is a 3D point in the MIS scene, then the 3D point can be transformed to the 2D point in the endoscopic camera's view by the following equation:

$$p_c = K * T_{cf} * P_f \tag{4.1}$$

where T_{cf} is the transformation from camera space to feature space (as shown in Figure 4.2 (a)) and can be computed by solving the Perspective-n-Point (PnP) problem, and K is the camera intrinsic parameters. For our proposed AR framework, as can be seen from Figure 4.2 (b), we add a surface local space S as an agent, which serves as the intermediary and is incrementally built from the point cloud sensed in the environment, which allows us to achieve great robustness. A 3D point in the model space P_m can be transformed to the 2D camera space p_c by:

$$p_c = K * T_{sc}{}' * T_{sm} * P_m \tag{4.2}$$

where T_{sc} is the transformation from surface space to camera space (as shown in Figure 4.2 (b)) and can be estimated by the SLAM. Here, the inverse of T_{sc} is used. And T_{sm} is a user-defined matrix that transform the space from surface to model. By using the local surface space as an agent, we solved two important issues for AR in MIS: (i) no pre-captured or manually selected features are needed, which saves time and enables 360 degree tracking; (ii) AR objects can be placed anywhere on the surface at the correct depth.

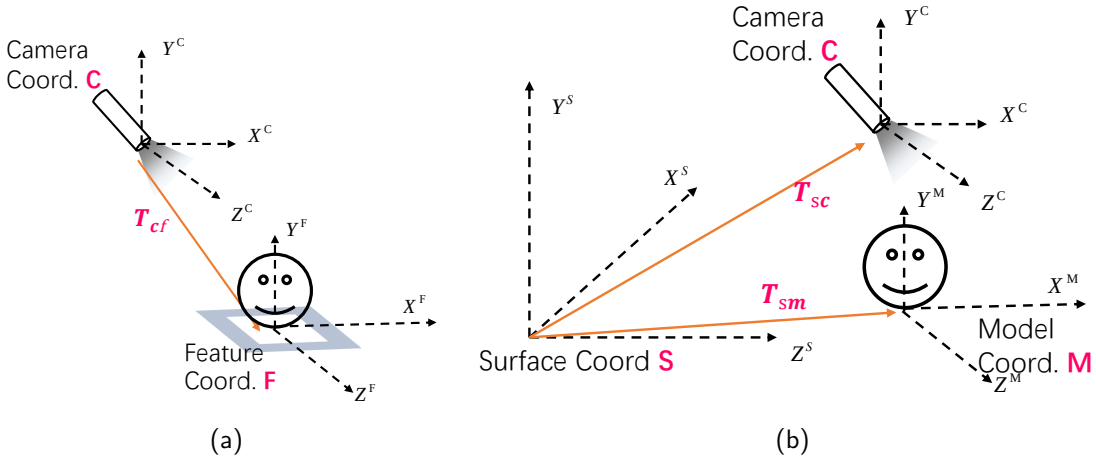


Figure 4.2: The comparison of the marker-less tracking (a) and our proposed AR framework (b).

4.2.2 Monocular Endoscopic Camera Tracking and Mapping

We use ORB–SLAM (Mur-Artal et al. 2015), which outperforms many SLAM systems such as Mono-SLAM (Davison et al. 2007), PTAM (Klein and Murray 2007) and LSD-SLAM (Engel et al. 2014), for the task of monocular endoscopic camera tracking and mapping. ORB–SLAM combines many state-of-the-art techniques into one SLAM system, such as using an ORB descriptor (Rublee et al. 2011) for tracking, local keyframe for mapping, graph-based optimization, the Bag of Words algorithm for re-localization, and an essential graph for loop closure.

ORB (Oriented FAST and Rotated BRIEF) descriptor (Rublee et al. 2011) is a binary feature point descriptor that calculate the intensity weighted patch located at keypoints and encode them into a 256-bit vector. It is an order of magnitude faster than SURF (Bay et al. 2006) and more than two orders faster than SIFT (Lowe 2004) with better accuracy. In addition, ORB features are invariant to rotation, illumination and scale, which means that it is capable of dealing with some of the main challenges in MIS scenes including rapid movements of endoscope cameras (rotation and zooming) and the change of brightness. ORB-SLAM has a keyframe selection mechanism that only keep non-redundant keyframes to reduce the computation of

bundle adjustment, followed by a graph optimization algorithm that optimize the all the poses of stored keyframe. ORB-SLAM has also embedded a bags of words place recognition module that performs loop detection to close possible loop and relocalize when the tracking is lost.

These features can enable real-time endoscopic camera tracking and sparse point mapping in an abdominal cavity as shown in Figure 4.1. Real-time performance is crucial in time-demanding medical interventions. Since ORB (Rublee et al. 2011) is a binary feature point descriptor, it is an order of magnitude faster than SURF (Bay et al. 2006) and more than two orders faster than SIFT (Lowe 2004) with better accuracy. In addition, ORB features are invariant to rotation, illumination and scale, which means that it is capable of dealing with some of the main challenges in MIS scenes including rapid movements of endoscope cameras (rotation and zooming) and the change of brightness.

4.2.2.1 Initialization

A common problem for monocular scene analysis using SLAM is the initialization, a step required for generating an initial map, because the depth cannot be recovered from a single image frame. An automatic approach is used in ORB–SLAM to calculate homography for planar scenes and a fundamental matrix for non-planar scenes dynamically. This approach can greatly increase the success rate of initialization and reduce the time required for the initialisation step. It also facilitates the initialization on an organ surface or to compute a fundamental matrix when the endoscopic camera is pointing at complex structures.

4.2.2.2 Training of Data Sets

One of the huge challenges that is unique to AR in MIS is the rapid movement of endoscopes due to constant extraction and insertion of the device. The tracking algorithm must be robust to accommodate the loss of image sequences after an extraction, and recover the tracking during a re-insertion. The Bags of Words (BoW) algorithm solves this re-localization problem during the tracking. In the BoW al-

Table 4.1: Comparison of the original and our trained BoW database

Item	***Original BoW Database***	***Database Trained for MIS***
Training Source:	Images with Different Genres	Endoscopy Videos
Database Size:	145.3 MB	41.8 MB
Number of Words:	971815	259677
BoW Query Time[1]:	4.85ms	4.42ms

gorithm, the vocabulary is created offline with a large number of ORB descriptors extracted from very large data sets of images that cover almost all of the patch patterns that may be encountered. The vocabulary serves as a classifier or a dictionary to assign each descriptor an index. When a new image appears in the system, each descriptor of features in this image is looked up, and a unique vector will be built based on the index of descriptors. In doing so, the rough similarity of two images can be acquired by simply comparing the two unique vectors, therefore, it can greatly increase the speed of re-localization.

The default BoW database in ORB–SLAM contains a very large image data set with different genres captured from the real world scenes. Such a universal database would be too sparse and general for specific MIS tasks. When processing endoscopic videos, images are generally captured inside of human bodies for different organs, tissues and vessels. These MIS scenes are more homogeneous and specific than the real word scenes. Therefore, we trained our vocabulary list specifically for its use in MIS based on 877 images sampled from ten *in vivo* sequences obtained from the Hamlyn Centre Endoscopic Video database (London 2016) (Ye et al. 2016). By training a specific MIS BoW database, the specific features existing in the minimally invasive surgery scenes are collected and saved. The length of the unique vectors for similarity measurements will be decreased, hence, reducing not only the loading time of the AR framework, but also the time of BoW query as shown in Table 4.1.

This approach generalizes well to different MIS scenes since the training based on the Hamlyn Centre Endoscopic Video Database covers a range of medical scenarios from gastrointestinal examinations, diaphragm dissection, lung lobectomy, coronary

[1]Based on the average time of 1000 times' BoW query experiment

artery bypass, to cardiac examination.

4.2.2.3 Parameter Tuning and Increasing Surface Points

We fine-tuned some of the parameters that were used by default in the ORB–SLAM by increasing the limit of the number of features extracted per image by a factor of two, which allows a maximum of 2000 feature points to be detected. The maximum threshold that is allowed between keypoints and reprojected map points for triangulation is reduced by a factor of ten to constrain the range of the points to be selected so that strictly robust 3D points are chosen and feature points moved by tissue deformation are rejected. This approach can greatly improve the tracking accuracy. Finally, the Hamming distance threshold for the ORB descriptor comparisons is decreased by 0.8 for more strict applications of the pair point rule. After tuning the default parameters, around 50% more map points can be detected for the reliable surface reconstruction pipeline. Furthermore, the system has the ability to filter small drifts caused by tissue deformations with strict map point selection criteria.

4.2.3 Intra-operative 3D Surface Reconstruction

One of the main advantages of our proposed AR system is its ability to use a sparse 3D point cloud extracted from a moving monocular endoscopic camera to construct a dense and smooth surface through our novel surface reconstruction framework. Our framework processes the unstructured sparse point clouds using a combination of outlier removal filters, the Moving Least Square algorithm to smooth noise data and a Poisson surface reconstruction method to generate the smooth surface from an unstructured sparse point cloud. This pipeline is illustrated in Figure 4.3. Details of each processing step are presented in the following sections.

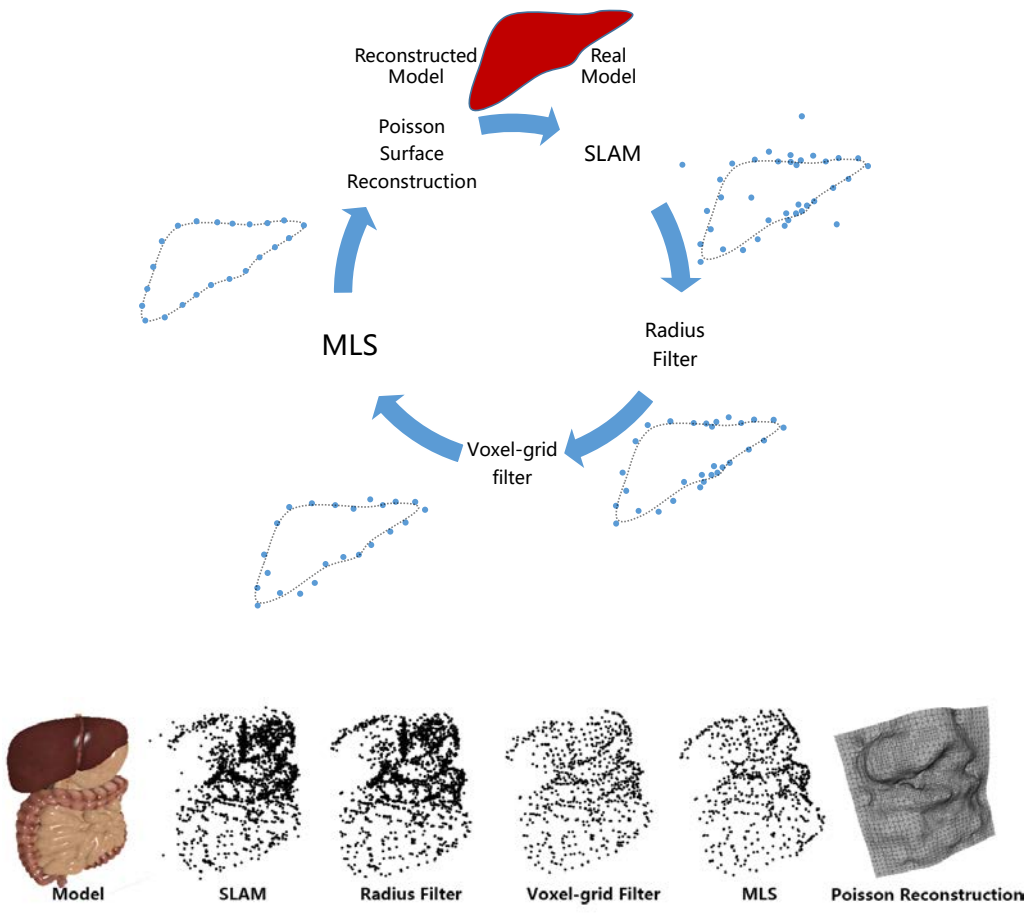


Figure 4.3: The proposed intra-operative 3D surface reconstruction framework.

4.2.3.1 Pointcloud Pre-processing

The point cloud P given by ORB–SLAM represents salient points that are visible at different camera keyframes, giving a sparse representation of the intra-operative scene. MIS scenes are very complex due to issues associated with camera calibrations and movements and reflections of tissues. Hence, the result is a noisy point cloud mixed with many outliers that can affect the final surface reconstruction. Our approach to solve this problem is to apply two outlier removal filters to remove the noisy points located amongst the raw data points before feeding the point cloud into the reconstruction pipeline.

Firstly, a radius filter is used to process points in a cloud based on the number of neighbour points. Points with very few neighbours are labelled as outliers or isolated points that should not contribute to the overall structure of the 3D scene. Since some texture-abundant areas gain many more points than other areas, a voxel-grid filter is then used to re-sample the point cloud into a more evenly distributed point cloud. After the filtering process, the point cloud becomes 'clean' and ready for MLS (Moving Least Square) smoothing and 3D surface reconstruction.

4.2.3.2 Moving Least Square Point Smoothing

The Moving Least Squares (MLS) algorithm (Levin 2004) reconstructs surfaces locally by solving an optimization problem to find a local reference plane and fit a polynomial to the surface. Let a point set $p_i \in \mathbb{R}_3, i \in \{1, ..., K\}$ be the point cloud produced from the ORB–SLAM system. the continuous and smooth MLS surface S can be computed by a two-step procedure: (i) a local reference plane is defined as $H = \{x \in \mathbb{R}_3 | x \cdot n - D = 0\}$, which can be computed by minimizing the weighted sum of squared distances:

$$\min_{<N,D>} \sum_{i=1}^{K} (p_i \cdot n - D)^2 \Phi(\|p_i - q\|) \tag{4.3}$$

where q is the projection of p onto H, and Φ is the MLS kernel, usually a Gaussian; (ii) after the points are projected onto the initial local reference plane, a second least squares optimization is used to find a bi–variate polynomial function $g(u, v)$ (where u, v is the local coordinate of q in H) that approximates to the local surface. The projection of p onto S can then be defined by the polynomial value at the origin, i.e. $q + g(0, 0) \cdot n$.

4.2.3.3 Poisson Surface Reconstruction

We represent the points after the MLS filter stage by a vector field $\overrightarrow{V}$, which is derived from N in previous section. Poisson surface reconstruction (Michael et al. 2006) approaches the surface reconstruction problem through a framework of implicit functions that compute a 3D indicator function χ (which is equal to 1 inside the model and 0 at the outside points). Therefore, the problem becomes that of finding the χ whose gradient is the best approximation of the vector filed $\overrightarrow{V}$:

$$min_{\chi} \left\| \bigtriangledown_{\chi} - \overrightarrow{V} \right\| \tag{4.4}$$

Applying the divergence operator, we can transform this into a Poisson problem:

$$\bigtriangledown \times (\bigtriangledown_{\chi}) = \bigtriangledown \times \overrightarrow{V} \Leftrightarrow \triangle_{\chi} = \bigtriangledown \times \overrightarrow{V} \tag{4.5}$$

After solving the Poisson problem and obtaining the 3D indicator function χ, the 3D surface can be directly obtained by extracting an isosurface (Kazhdan and Hoppe 2013). The Poisson reconstruction process acts as a global solution that treats all of the data points simultaneously without relying on a heuristic partitioning or blending, so that it can robustly approximate noisy data and create very smooth surfaces.

4.3 Results

We designed a two-part quantitative and qualitative evaluation process: (i) using a realistic simulation of a MIS scene video for the ground truth study to assess the performance of the SLAM tracking error and the accuracy of the proposed surface reconstruction framework; (ii) using a real *in vivo* video acquired from the Hamlyn Centre Laparoscopic/Endoscopic Video Datasets (London 2016) (Mountney and Yang 2010b) to assess the quality of our proposed framework.

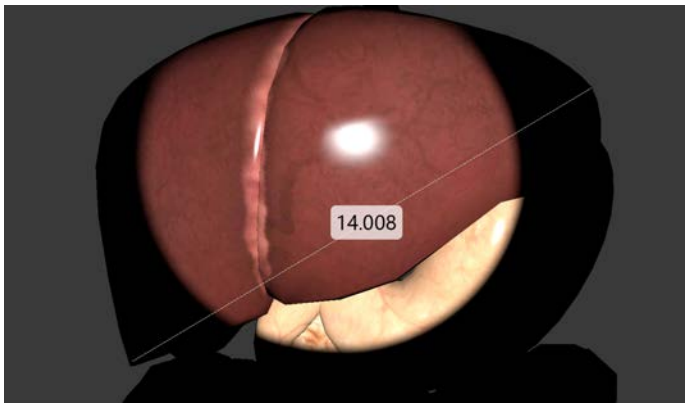
4.3.1 System Setup

Our system is implemented in an Ubuntu 14.04 environment using C/C++ (without any GPU acceleration). All experiments are conducted on a workstation equipped with Intel Xeon(R) 2.8 GHz quad core CPU, 32G Memory, and one NVIDIA GeForce GTX 970 graphics card. The size of the simulation image sequences is 1024 X 768 pixels and the size of *in vivo* endoscope video is 840 X 640 pixels. ORB–SLAM with our proposed AR framework runs in real-time at 40 FPS at max and the 3D surface reconstruction process takes around 600ms to traverse the whole pipeline (which is only trigger once, and not calculated for every frame).

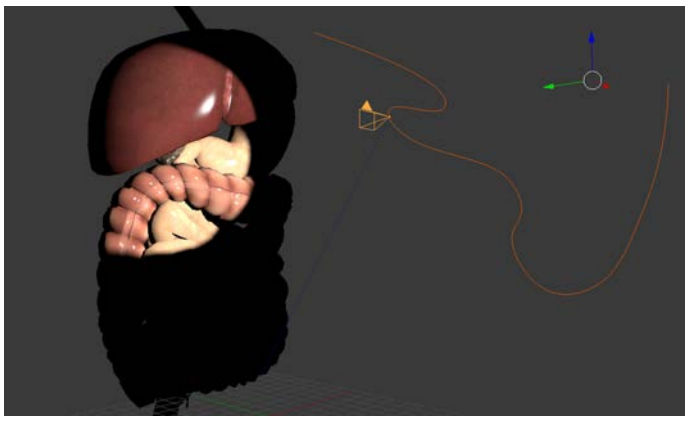
4.3.2 Ground Truth Study using Simulation Data

For the evaluation of the accuracy of tracking performance, all camera trajectories estimated by ORB–SLAM were aligned with trajectories of the ground truth camera used to render the MIS scene video. Similarly, the accuracy of our proposed 3D surface reconstruction framework is evaluated by comparing the reconstructed surface with the 3D model used to render the simulation video.

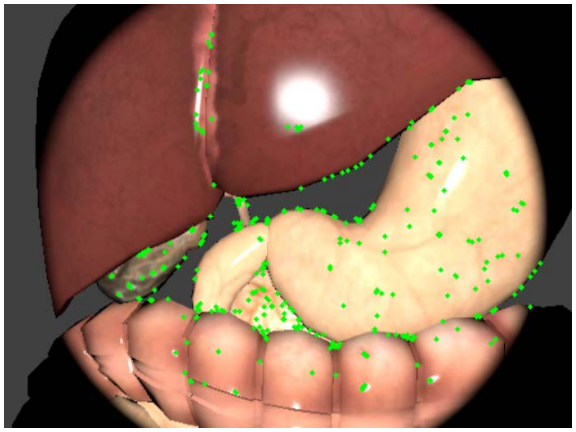
To quantitatively evaluate the performance of ORB–SLAM, we used Blender (Blender 2016) – an open source 3D creation software to render realistic image sequences of a simulated abdominal cavity scene using pre-defined endoscopic camera movements. The digestive system contains 3D models with textures to make the scene as realistic as possible. The model was scaled to be the real life size according



(a)



(b)



(c)

Figure 4.4: Simulated MIS scenes with a realistic human digestive system model. (a) The size of the model is scaled to the real world size of an adult liver. (b) The only light is attached to the camera and the camera trajectory is designed to hover around the 3D model. (c) The frame that ORB–SLAM succeeded in initializing.

to an average measured liver diameter of 14.0 cm (Kratzer et al. 2003) as shown in Figure 4.4(a), the material property was set with a strong specular component to simulate the smooth and reflective liver surface tissue. The luminance is intentionally set high with a spot light attached to the main camera to simulate an endoscope camera as shown in Figure 4.4(a) to render a realistic endoscopic lighting condition. We designed a camera trajectory that hovers around the 3D model (Figure 4.4(b)) to capture as much of the area as possible so as to build a point cloud that could cover the whole front surface of the models. Nine hundred frames of image sequences were captured at a frame-rate of 30 fps, which is equivalent to a 30 second video. In order to investigate the robustness of our framework, we intentionally add white noise with different standard deviation (SD) to the synthetic video. We now have three version of the synthetic videos (with no white noise, white noise SD=1, and white noise SD=3, respectively), which will together be used for the further evaluation.

4.3.2.1 Camera Trajectory Evaluation

Figure 4.4(c) shows one of the rendered images from the sequences used as the input to ORB–SLAM. The camera trajectory started with a close shot location of the liver surface. ORB–SLAM was successfully initialized around frame 200 to 300 when the camera was in a place and where many feature points were identified. After the initialization step, the SLAM system ran stably and the camera trajectory was estimated with the origin of the coordinate system at the initialized position. The estimated camera trajectory was then extracted and normalized into the same coordinate system as that of the simulated ground truth model to assess the SLAM tracking performance.

Figure 4.5 shows the performance evaluation results; Figure 4.5(a) displays the camera trajectories in 3D space, in which green, dark blue and light blue dots represent the camera trajectory estimated by ORB–SLAM under no white noise, white noise SD=1 and white noise SD=3. Red dots are the trajectory of the simulated ground truth. Figs. 4.5(b), (c), and (d) shows the camera trajectories in X-axis, Y-axis, and Z-axis views, respectively. As can be seen, the SLAM camera trajectory

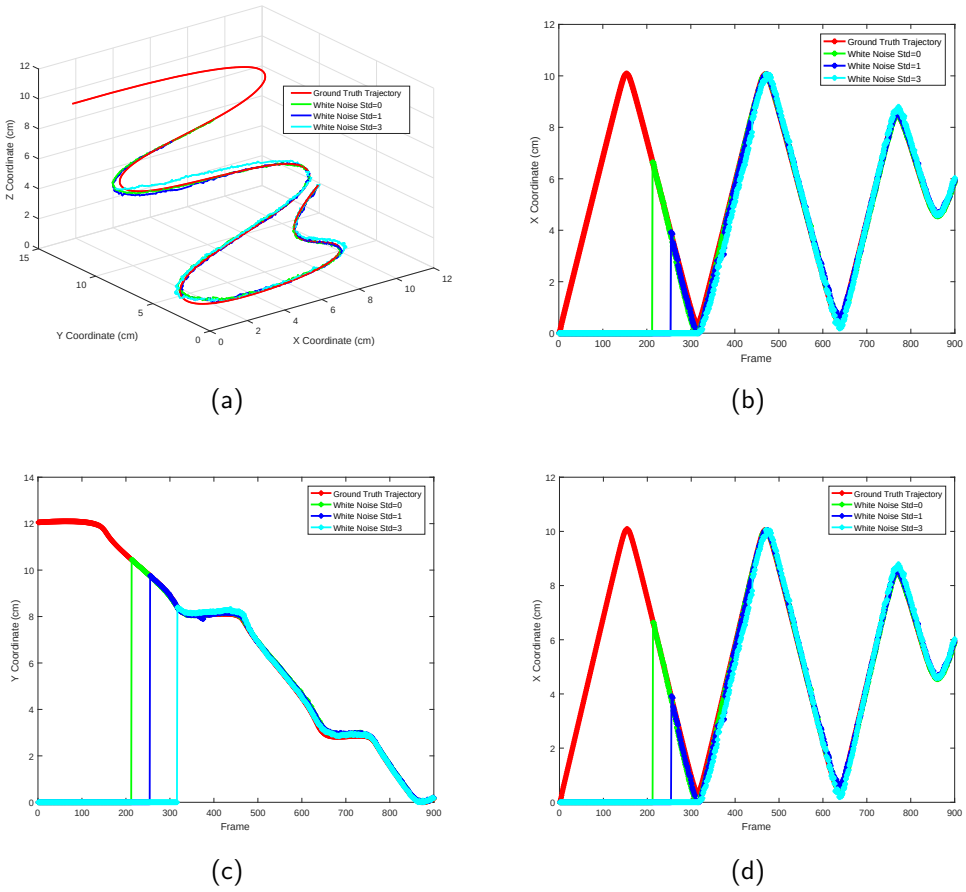


Figure 4.5: The camera trajectory comparison of the ground truth (red dots) with the estimated results under different white noise levels: no white noise (green dots), white noise SD=1 (dark blue dots), and white noise SD=3(light blue dots) in four different views, (a) 3D view, (b) view of X-axis, (c) view of Y-axis, (d) view of Z-axis

starts at frame 212, 254 and 316 for the video with no white noise, white noise SD=1 and white noise SD=3, respectively, as there is no estimated data before initialization. Once the camera tracking is initialized, the trajectory of the camera matches closely with the ground truth camera trajectory represented by red dots. RMSE between the two camera trajectory data sets was also calculated with results of 1.24mm, 2.33mm and 4.39mm.

4.3.2.2 3D Surface Reconstruction Evaluation

When the ORB–SLAM system gained enough feature points, we build a 3D surface based on the sparse point cloud. The whole reconstruction pipeline takes only 600 ms to generate the surface, which was then exported into the 3D model space to be compared with the ground truth surface data set. A simple iterative closest point (ICP) algorithm was used to align the reconstructed surface with the 3D model that was used to render the video. Root Mean Square Distance (RMSD) is used to evaluate the overall distance between the two surfaces. They are aligned in the real world coordinate system and we apply a grid sample to get a series of x,y coordinate points based on the surface area, and then compare the distance of the z value of the two surfaces.

$$RMSD = \sqrt{\frac{1}{mn}\sum_{x=1}^{m}\sum_{y=1}^{n}(Z_{x,y} - z_{x,y})^2}$$

The RMSD to the ground truth surface is 2.54mm, 2.81mm and 3.66mm for the surface reconstructed by our proposed framework with different white noise levels. As shown in Table. 4.2, our proposed method is much more accurate than the Shape-from-Shading (SFS) (Prados and Faugeras 2005)(Visentini-Scarzanella et al. 2012) method as SFS was based on the strong assumption of a single point illumination source and can be affected by different tissue colours. Also, the reconstruction error of our method is better than that of the sparse cloud points reported as 4.10mm (Mahmoud et al. 2016)(Mur-Artal et al. 2015). Also, our method can reconstruct a dense surface compared to that of the less clinical applicable sparse method. To further evaluate our reconstruction result, we have also rendered our video in stereo

Table 4.2: Surface reconstruction results

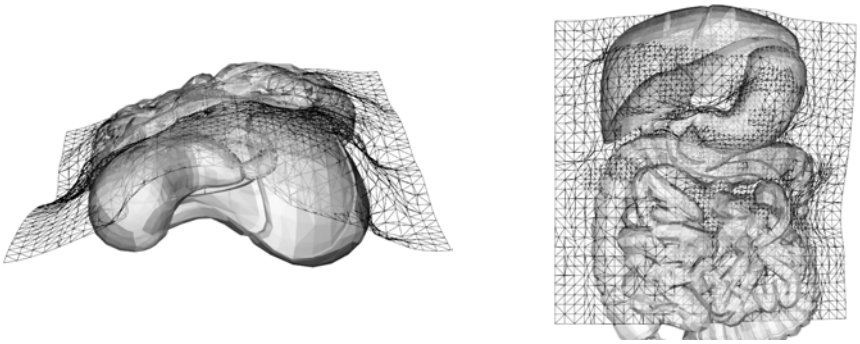
Type	Method	RMSD(wn=0)	RMSD(wn=1)	RMSD(wn=3)
Mono/Dense	SFS(Visentini-Scarzanella et al. 2012)	7.21mm	8.38mm	11.60mm
Mono/Dense	**Proposed**	**2.54mm**	**2.81mm**	**3.66mm**
Stereo/Dense	BM (Chen et al. 2001)	2.04mm	2.09mm	2.17mm
Stereo/Dense	Chang *et al*(Chang et al. 2013)	2.57mm	2.21mm	2.28mm

mode and tested it with popular stereo reconstruction approaches such as Block Matching (BM) and the state-of-the-art cost volume stereo reconstruction method by Chang *et al*. Our method is slightly better than the cost volume when there is no white noise, but overall is less accurate than stereo reconstruction, as the depth can be directly calculated from the disparities of stereo image pairs.

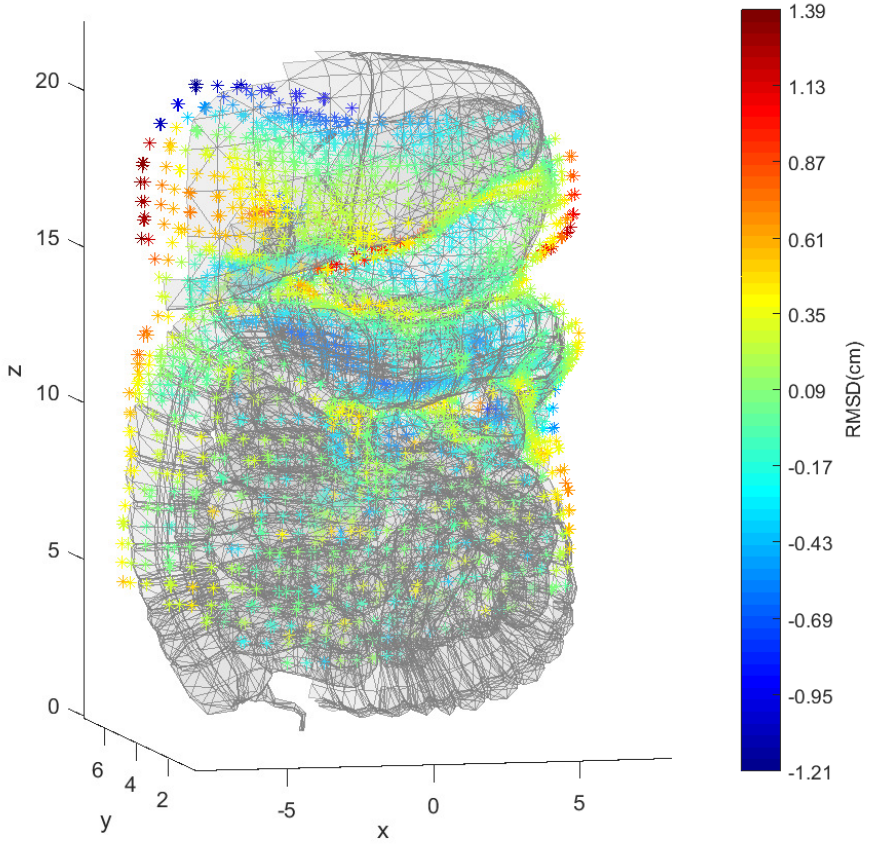
Figure 4.6 (a) shows that the reconstructed 3D surface aligns with the 3D model closely; Figure 4.6 (b) shows the top down view of the alignment. Figure 4.6 (c) shows the measured points between the reconstructed surface with the 3D ground truth model to illustrate the position of the measured points, where warm colours show penetrations between the two surfaces, the green colour represents a perfect match between the two surfaces, and the blue colour shows the largest distance between the two surfaces.

4.3.3 Real Endoscopic Video Evaluation

To qualitatively evaluate the performance of our proposed surface reconstruction framework, we applied the proposed approach with the real *in vivo* videos from the Hamlyn Centre Laparoscopic / Endoscopic Video Datasets. Figure 4.7 (a) (e) and (f) shows the reconstruction results from our 3D reconstruction framework. Figure 4.7 (b) shows the depth augmentation by fusing the camera pose from the SLAM system and the 3D surface reconstructed from our proposed framework. The real-time alignment of the 3D transparent mesh and the video are a good match that the mesh is closely matched with the 3D model, suggesting that our method can provide the correct depth information intra-operatively and so help improve surgical performance by displaying 3D mesh structures when performing monocular endoscope procedures. However, when large deformation occurs or the surgical



(a) (b)



(c)

Figure 4.6: (a) and (b): the surface nicely represents the model surface. (c) Surface reconstruction error map.

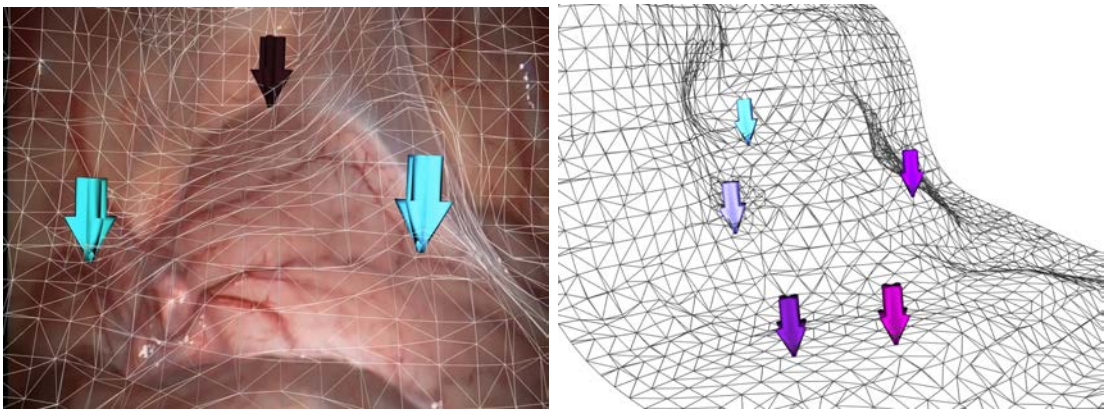
instruments occupy the large proportion of the view, our framework may fail as shown in Figure 4.7 (f).

With our new 3D surface reconstruction approach, we have developed a geometry-aware AR framework for depth correct AR argumentation within the intra-operative endoscope scene in real-time. Our AR framework is an important step towards high quality AR in MIS, since incorrect depth placement will cause virtual objects to appear to drift away when the viewing angle changes. Furthermore, accurate global geometric information plays a crucial role in augmenting the real surgical scenes with annotations, labels, tumour measurements, inguinal measurements to estimate optimal mesh size for inguinal herniorrhaph (Knook et al. 2001) or even a 3D reconstruction of anatomy structures at the target surgical location. We demonstrate two example applications to show the clinical relevance.

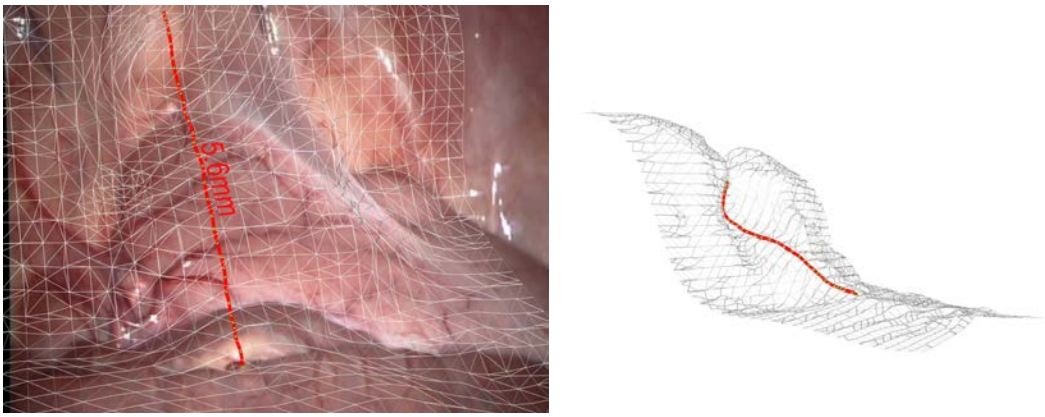
In Figure4.7 (a), AR augmentations of 3D arrows labels are placed onto the video frames to generate artificial depth cues and Figure4.7 (b) shows that virtual 3D arrows exist at different depths within this geometry-aware environment. In the second example, we recover the scale to the real-world size (Nützi et al. 2011) to enable accurate intra-operative measurement as demonstrated in Figure4.7 (c) and (d). Note that measurement (the red line) follows the surface curvature closely, providing accurate results with correct depth information. More details can be appreciated in our demonstration video (Chen 2016).

4.4 Discussion

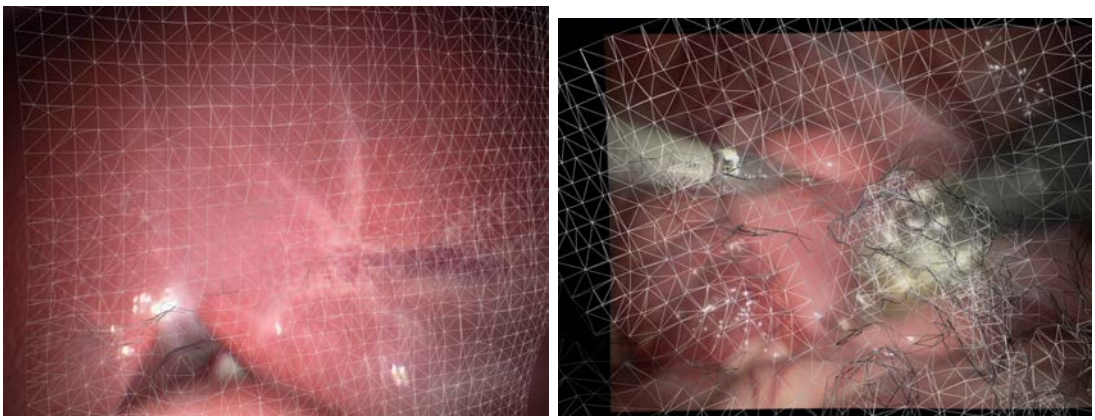
Intra-operative MIS scene reconstruction is a challenging task especially for monocular MIS scene that the only input source is the monocular video stream. Acquiring the depth and geometric information in MIS is crucial for not only AR tasks such as intra-operative measurement, but also enables the potential applications of skill evaluation (Jiang et al. 2017), autonomous tasks such as autonomous ultrasound scanning (Zhang et al. 2017), debridement and cutting (Murali et al. 2015). We are able to achieve a promising reconstruction result by our proposed SLAM-based



(a) (b)



(c) (d)



(e) (f)

Figure 4.7: The surface reconstruction results applied to an *in vivo* video sequence. (a) Interactively adding arrows as annotations intra-operatively.(b) The view of mesh to show the annotations are in different depth.(c) Intra-operative measurement example. (d)The side-view of the intra-operative measurement example. Note that the measurement line follows the surface curvature closely. (e) The augmented mesh on a liver. (f) Our framework may fail when large deformation occurs or the surgical instruments occupy large proportion of the view

monocular reconstruction approach (RMSD = 2.54mm), which is much accurate than other monocular MIS scene reconstruction method (RMSD = 7.21mm) and even comparable to the state-of-the-art stereo reconstruction method (RMSD = 2.04/2.57mm) that the depth can be directly derived from the disparity of stereo vision.

Using simulation for quantitative evaluation is a novel part of this Chapter, including the details of how to render realistic video and scale the model for accessing real-world size of accuracy. In this way, the accuracy can be obtained at no cost and without the need of real experiment (which is very expensive especially for medical surgery). However, although for tracking and reconstruction tasks, the realistic texture doesn't affect much, the evaluation on simulation sometimes still cannot represent the accuracy in real-world. As in real world scenario, there are tissue motion, movement of instrument and blood, smoke that will affect the tracking and reconstruction results, which is usually hard to simulate.

The limitation of our proposed method is that the SLAM theory is developed based on static world assumption; the deformations of objects (such as tissues and organs) directly challenge this basic condition for SLAM to estimate camera poses for 3D reconstruction. Therefore, soft tissue deformation is a great challenge to support in the SLAM based reconstruction framework as proposed here. Particularly with monocular endoscopic videos, it is extremely hard to recovery the soft deformation correctly while simultaneously estimating the camera poses. For small deformations like those in the *in-vivo* video that we use, however, the RANSAC algorithm in SLAM system will filter the outliers and recover the correct movement. For large deformation in very small FOVs, it is still unclear how to solve the tissue deformation issue without using extra external sensors within the monocular scene.

Using stereoscopic views is a possibility and we will investigate this in future work. One possible solution to accurately simulate and track the deformation is to to use real-time deformation model(Zou and Liu 2017) and feature-based tracking (Kumar et al. 2014) to recovery the movement of tissue. Although the accuracy and speed of our framework are acceptable, we will continue developing a dense SLAM

system to be used in MIS reconstruction and extend the current reconstruction framework. This will enable us to develop a prototype system that can be tested in the operating theatre with our clinical collaborators, further investigating the benefit and efficacy of our approach and providing evidence for our hypothesis that visual SLAM can enhance the tools available to surgeons performing monocular endoscopic procedures.

4.5 Conclusions

In this chapter, we presented an efficient and effective 3D surface reconstruction framework for an intra-operative monocular laparoscopic scene based on SLAM. This new approach has shown promising results when tested on both simulated laparoscopic scene image sequences and clinical data. The proposed framework also reveals several potential clinical applications such as additional depth cues augmentation and geometry-aware augmented reality in MIS.

Chapter 5

Stereo-based Online Global Surface Reconstruction for GA-AR

5.1 Introduction

Laparoscopic surgery is a Minimally Invasive Surgical (MIS) procedure using endoscopes with small incisions to carry out internal operations on patients. While MIS offers considerable advantages over open surgeries, it also imposes big challenges on surgeons' performance due to the well known MIS issues associated with the field of view (FOV), hand-eye dis-alignment and disorientation. Augmented Reality(AR) technology can help overcome the limitations by overlaying additional information with the real scene through augmentation of target surgical locations, annotations (Kim et al. 2012), labels (Su et al. 2009), tumour measurement (Bourdel et al. 2017) or even 3D reconstruction of anatomic structures (Haouchine et al. 2013) (Haouchine et al. 2015).

The potential of Augmented Reality (AR) technology to assist minimally invasive surgeries (MIS) lies in its computational performance and accuracy in dealing with challenging MIS scenes. Even with the latest hardware and software technolo-

gies, achieving both real-time and accurate augmented information overlay in MIS is still a formidable task. Despite recent advances in powerful miniaturised AR hardware devices and improvements on vision based software algorithms, many issues in medical AR remain unsolved. Especially, the dramatic changes in tissue surface illumination and tissue deformation as well as the rapid movements of the endoscope during the insertion and the extrusion all give rise to a set of unique challenges that call for innovative approaches.

As with in any technological assisted medical procedures, accuracy of AR in MIS is paramount. With the use of traditional 2D feature based tracking algorithms such as those used in (Du et al. 2015) ,(Planteféve et al. 2016), (Kim et al. 2012) and (Mountney and Yang 2008), the rapid endoscope movement can easily cause feature points extracted from the vision algorithms to fall out of the field of view, resulting in poor quality visual guidance. The latest visual SLAM (Simultaneous Location and Mapping) based approaches have the potential to overcome this issue by building an entire 3D map of the internal cavity of the MIS environment, but SLAM algorithms are often not robust enough when dealing with tissue deformations and scene illuminations (Turan et al. 2017) (Klein and Murray 2007) (Klein and Murray 2007) (Mur-Artal et al. 2015) (Mahmoud et al. 2016). Furthermore, in order to meet the demand of high computational performance, sparse landmark points are often used in MIS AR, and augmented information are mapped using planar detection algorithms such as Random Sample Consensus (RANSAC) (Lin et al. 2013) (Grasa et al. 2014). As a result, AR content is mapped onto planes rather than curved organ surfaces.

In this chapter, we present a novel real-time AR framework for MIS that achieves interactive geometric aware augmented reality in endoscopic surgery with stereo views. Our framework tracks the movement of the endoscopic camera and simultaneously reconstructs a dense geometric mesh of the MIS scene. The movement of the camera is predicted by minimising the re-projection error to achieve a fast tracking performance, while the 3D mesh is incrementally built by a dense zero mean normalised cross correlation (ZNCC) stereo matching method to improve the

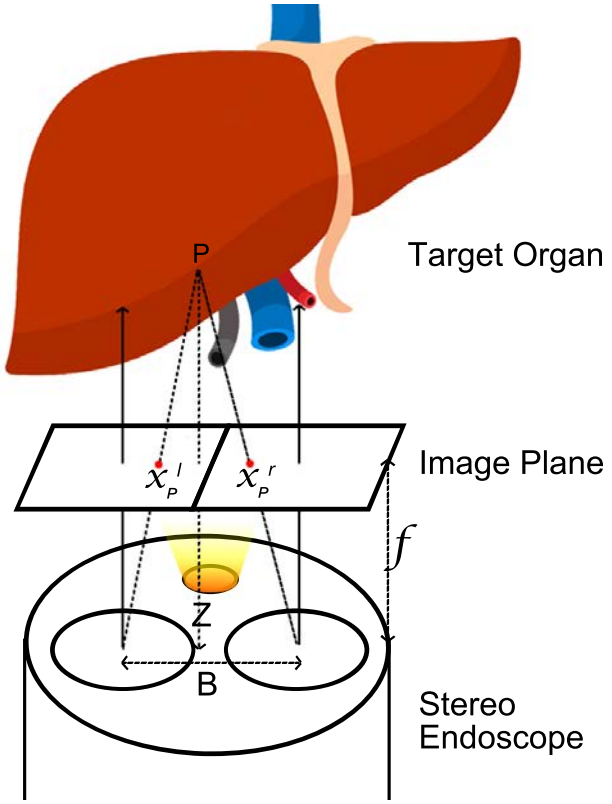


Figure 5.1: By using a stereo endoscope, the 3D position of any point in the view can be directly estimated by using stereo triangulation.

accuracy of the surface reconstruction. Our proposed system does not require any prior template or pre-operative scan and can infer the geometric information intra-operatively in real-time. With the geometric information available, our proposed AR framework is able to interactively add annotations, localisation of tumours and vessels, and measurement labelling with greater precision and accuracy compared with the state of the art approaches.

5.2 Methods

As can be seen from the flowchart in Figure 5.2, our proposed framework starts with a SLAM system that can track and estimate the camera pose frame by frame. The following stereo matching algorithm based on ZNCC is used to reconstruct dense surface at each keyframe, which is then transformed and stitched to a global surface

based on the inverse transformation of the camera pose. Finally, the global surface is re-projected to 2D based on the camera pose and overlaid on the image frame, serving as an interactive geometric layer. The geometric layer enables the interactive AR applications such as online measurement as shown later in Figure 5.4.

5.2.1 Landmark Point Detection and Triangulation

In medical interventions, real-time performance and accuracy are both critical. We adopt Oriented FAST and Rotated BRIEF (ORB) (Rublee et al. 2011) feature descriptors for feature points extraction, encoding and comparison to match landmark points in left and right stereo images. ORB is a binary feature point descriptor that is an order of magnitude faster than SURF (Bay et al. 2006), more than two orders faster than SIFT (Lowe 2004) and also offers better accuracy than SURF and SIFT (Rublee et al. 2011). In addition, ORB features are invariant to rotation, illumination and scale, hence, capable of dealing with challenge endoscope camera scenes (rapid rotating, zooming and changing of brightness).

We apply the ORB detector and find the matched keypoints on left and right images. Let x_P^l and x_P^r be the x coordinates on the left and right images, respectively. Assuming the left image and the right image are already rectified, the focal length of both cameras f and the baseline B are known fixed values, by similar triangles, the depth or the perpendicular distance Z between the points and the endoscope can be found according to similar triangles (see Figure 5.1).

$$\frac{B - \left(x_P^l - x_P^r\right)}{Z - f} = \frac{B}{Z} \quad \Rightarrow \quad Z = \frac{f \cdot B}{d_P} \tag{5.1}$$

Where $x_P^l - x_P^r$ is the disparity d_P of the two corresponding keypoints in the left and the right images detected by the ORB feature.

We then perform a specular reflection detection by removing the keypoints that have intensities above a threshold for efficiency. This could effectively remove the influence of specular reflections from the next stage of computation.

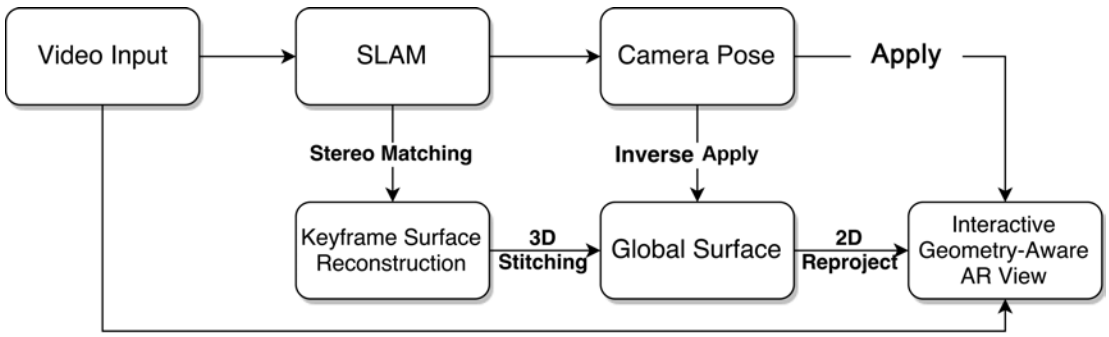


Figure 5.2: Flowchart describing the whole framework

5.2.2 Frame by Frame Camera Pose Estimation

Any AR application requires the real-time frame by frame tracking to continuously update the overlay positions. To meet the real-time requirement, after initialization, we employ the constant velocity motion model used by MonoSLAM (Davison et al. 2007) to roughly estimate the position r_{t+1} and quaternion rotation q_{t+1} of the camera position based on the current linear velocity v_t and angular velocity w_t in a small period Δt:

$$\left.\begin{aligned} r_{t+1} &= r_t + v_t \cdot \Delta t \\ q_{t+1} &= q_t \times q\left(\omega_t \cdot \Delta t\right) \\ v_t &= v_{t-1} + a_{t-1} \cdot \Delta t \\ \omega_t &= \omega_{t-1} + \alpha_{t-1} \cdot \Delta t \end{aligned}\right\} \tag{5.2}$$

Based on the predicted camera pose (r_{t+1}, q_{t+1}), the potential regions where the feature points may appear on the image are estimate by re-projection of 3D points, hence reducing searching areas and computational cost.

A RANSAC procedure is then performed to obtain the rotation and translation estimations from the set of all the inlier points. During each RANSAC iteration, 3 pairs of corresponding 3D points from current point set p_t^i and point set in next period p_{t+1}^i are selected randomly to calculate the rotation matrix R and the trans-

lation t, which minimizes the following objective function:

$$\min_{<R,T>} \sum_{i=1}^{n} \left\| p_t^i - \left(R * p_{t+1}^i + T \right) \right\| \tag{5.3}$$

From the set with smallest re-projection error, the set of outlier points is rejected and all the inliers are used for a refinement of the final rotation and translation estimations.

During the inlier/outlier identification scheme by RANSAC (Fischler and Bolles 1981a) (an iterative method to estimate parameters of a mathematical model from a set of observed data that contains outliers), false matched ORB feature points, moving specular reflection points and deforming points are effectively rejected. This is a very important step for a MIS scene where the tissue deformation caused by respiration and heartbeat, as well as blood, smoke and surgical instruments can have impact on the tracking stability. Therefore, at this stage, we use the strategy to filter out any influence caused by occlusion and deformation to recover the camera pose. Indeed, the deformable surface is an unsolved challenge in MIS AR, we address this issue by reconstructing a dense 3D map through a more efficient stereo matching method (see Section 5.2.4).

5.2.3 Keyframe-based Bundle Adjustment

As our camera pose estimation is only based on the last state, the accumulation of error over time would cause system drifting. However, we cannot perform a global optimization for every frame as this will slow down the system over time. We follow the successful approach of PTAM (Klein and Murray 2007) and ORBSLAM (Mur-Artal et al. 2015) in correcting system drafting, which use the keyframe-based SLAM framework to save "snapshots" of some frames as keyframes to enhance the robustness of the tracking whilst not increasing computational load on the system. Each keyframe is selected based on the criteria that the common keypoints of the two keyframes are less than 80% keypoints but the total number exceeds 50.

Once a keyframe is assigned, bundle adjustment (BA) is applied to refine the 3D

positions of each stored keyframe KF_i and the landmark points P_j by minimising the total Huber robust cost function ρ_h – the re-projection error between 2D matched keypoints $p_{i,j}$ and camera perspective projections of the 3D positions of keyframes KF_i and the landmark points P_j:

$$arg \min_{KF_i, P_j} \sum_{i,j} \rho_h \left(\| p_{i,j} - CamProj \left(KF_i, P_j \right) \| \right) \tag{5.4}$$

5.2.4 ZNCC Dense Stereo Matching

We create a feature-based visual odometry system for the endoscopic camera tracking and landmark points mapping, which takes into account of illumination changes, specular reflections and tissue deformations in MIS scenes, as they usually appear as outliers, which can be effectively removed by the RANSAC algorithm. However, as the sparse landmark points can barely describe the challenging environment of MIS scenes, we perform a dense stereo matching upon the landmark points to create a dense reconstruction result.

The dissimilarity measure used during the stereo matching is a patch-based ZNCC method. The cost value $C(p, d)$ for a pixel p at disparity d is derived by measuring the ZNCC of the pixel in the left image and the corresponding the pixel $p - d$ in the right image:

$$C\left(p, d\right) = \frac{\sum_{q \in N_p} \left(I_L(q) - \bar{I}_L(p) \right) \cdot \left(I_R(q-d) - \bar{I}_R(p-d) \right)}{\sqrt{\sum_{q \in N_p} \left(I_L(q) - \bar{I}_L(p) \right)^2 \cdot \sum_{q \in N_p} \left(I_R(q-d) - \bar{I}_R(p-d) \right)^2}} \tag{5.5}$$

where $\bar{I}\left(p\right) = \frac{1}{N_p} \sum_{q \in N_p} I\left(q\right)$ is the mean intensity of the patch N_p centered at p.

ZNCC is proven to be less sensitive to illumination changes and can be parallelised efficiently on a GPU (Stoyanov et al. 2010). A WTA (Winner-Takes-All) strategy is applied to choose the best disparity value for each pixel p, followed by a convex optimization to solve the cost volume constructed by Huber-L^1 variational energy function (Chang et al. 2013) for a smooth disparity map. We used the GPU

implement of ZNCC and convex optimization for the efficient disparity map estimation and filtering in real-time.

5.2.5 Incremental Building of Geometric Mesh

The 3D dense points estimated by stereo matching are transformed to the world coordinate system by the transformation matrix from frame space to the world space T_{f2w} that was estimated by our feature-based SLAM system. A fast triangulation method (Marton et al. 2009) is then used to incrementally reconstruct the dense points into a surface mesh. Fig. 5.3 demonstrates the incrementally building process from Frame 1 to 900. The first and third rows are the reconstructed geometric mesh while the second and fourth rows are the current video frames. The geometric mesh can be built incrementally to form a global mesh that can then be re-projected back to the camera's view using the estimated camera pose for the augmented view (see (a) and (c) in Fig. 5.4).

5.3 Results and Discussion

We have designed a two-parts assessment process to evaluate our AR framework: (i) using a realistic 3D simulated MIS scene as the ground truth study to measure the reconstruction error by measuring the difference between the ground truth values and the reconstructed values; (ii) using a real *in vivo* video acquired from the Hamlyn Centre Laparoscopic/Endoscopic Video Datasets (London 2016) (Mountney and Yang 2010b) to assess the quality of applications of our proposed framework i.e. measurements, adding AR labels and areas highlighting.

5.3.1 System setup

Our system is implemented in an Ubuntu 14.04 environment using C/C++. All experiments are conducted on a workstation equipped with Intel Xeon(R) 2.8 GHz quad core CPU, 32G Memory, and one NVIDIA GeForce GTX 970 graphics card.

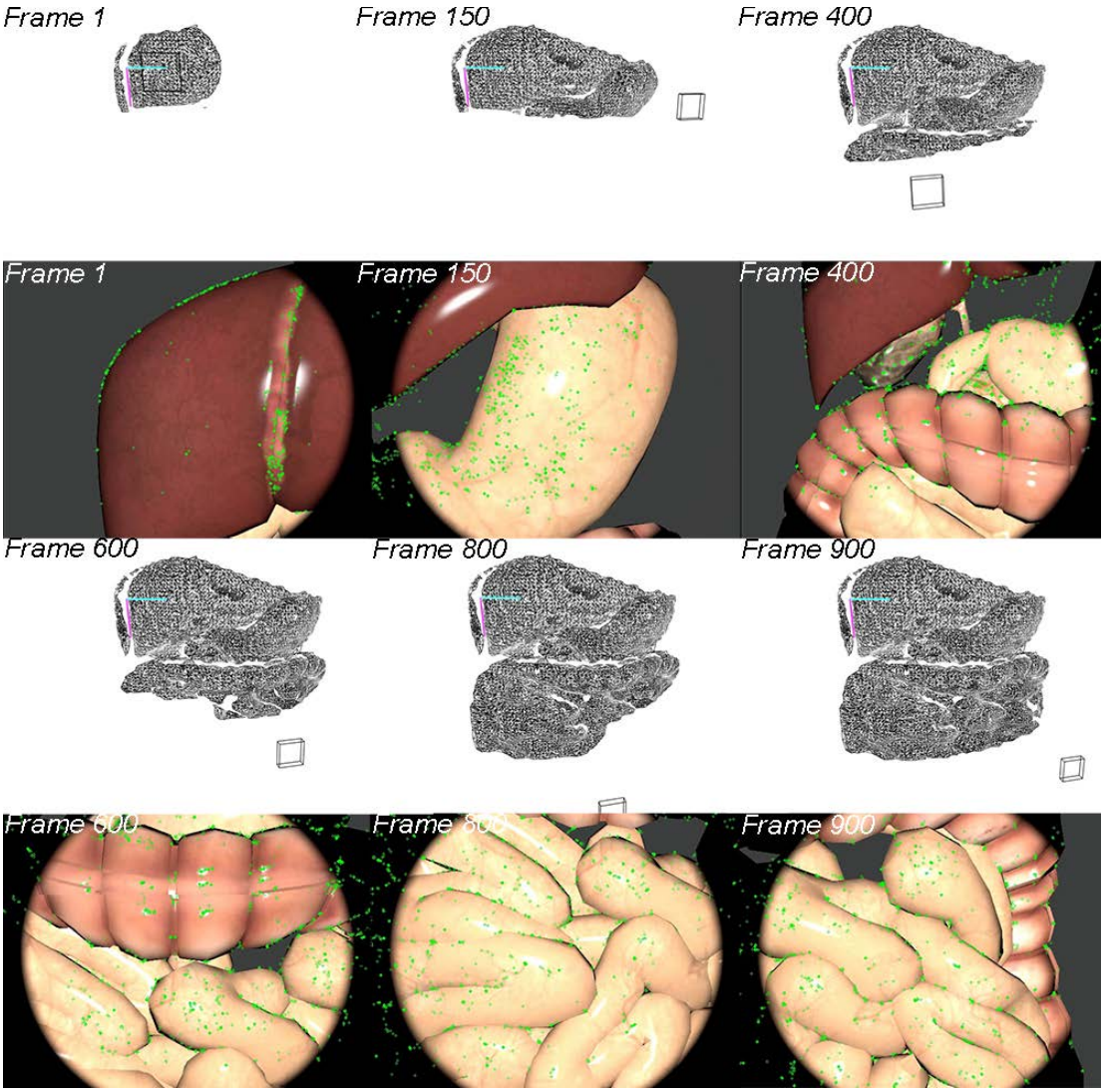


Figure 5.3: Incrementally building the geometric mesh. Rectangular boxes are the estimated camera pose; Green points are detected landmark points.

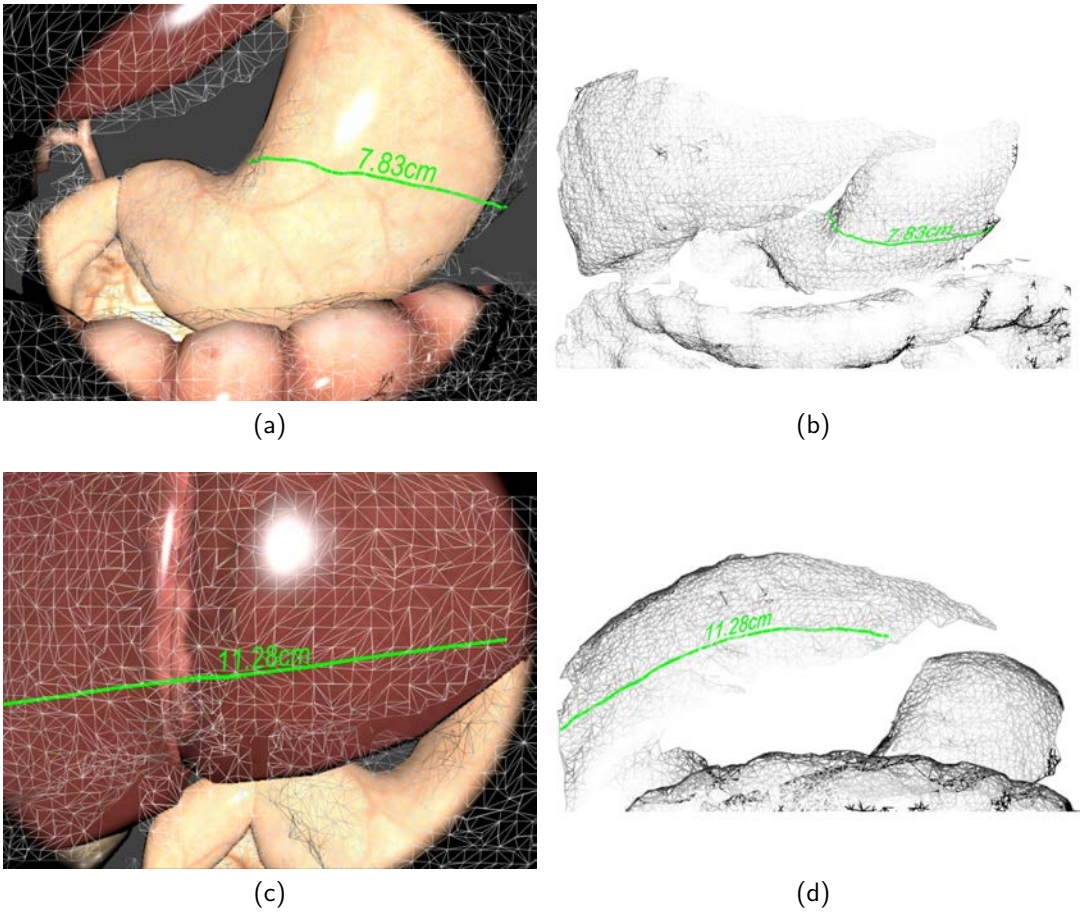


Figure 5.4: Measurement application of our proposed geometry-aware AR framework. Note that the measuring lines (green lines) accurately follow along the curve surface.

The size of the simulation image sequences and *in vivo* endoscope videos is 840 X 640 pixels. The AR framework and 3D surface reconstruction run in different threads. The 3D surface reconstruction process takes about 200ms to traverse the entire pipeline for each frame. Our proposed AR framework can run in real-time at 26 FPS when the reconstruction only performs at keyframes.

5.3.2 Ground Truth Study using Simulation Data

The performance of our proposed framework is measured in terms of reconstruction accuracy by comparing the reconstructed surface with the 3D model used to render the simulation video.

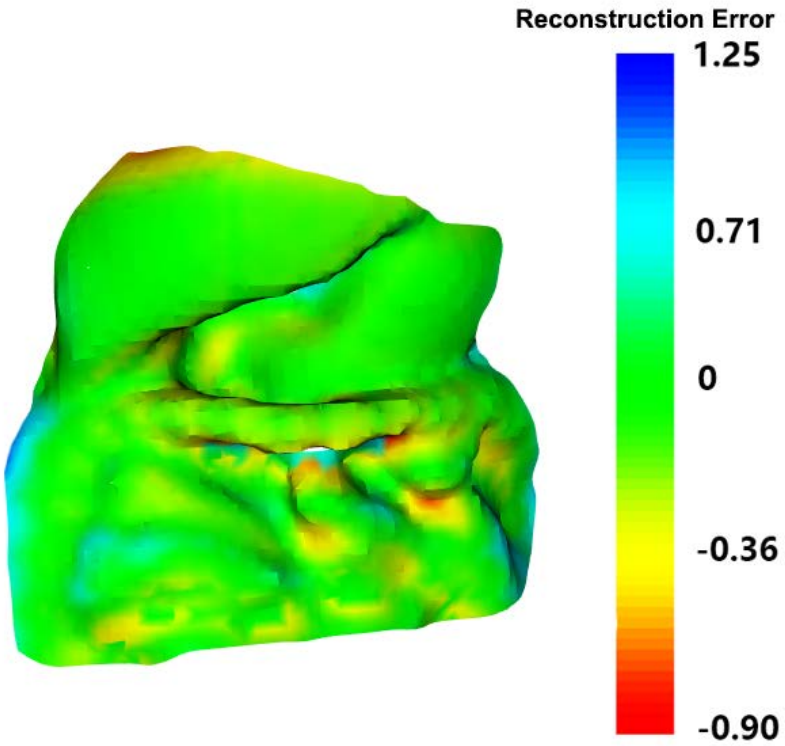


Figure 5.5: Reconstruction error map.

To quantitatively evaluate the performance of the progressive reconstruction result, we used Blender (Blender 2016) – an open source 3D software to render realistic image sequences of a simulated abdominal cavity scene using a set of pre-defined endoscopic camera movements. The simulated scene contains models scaled to real life size according to an average measured liver diameter of 14.0 cm (Kratzer et al. 2003), and the digestive system is rendered with appropriate textures to make the scene as realistic as possible. The material property is set with a strong specularity component to simulate the smooth and reflective liver surface tissue. The luminance is intentionally set high to simulate an endoscope camera as shown in Fig. 5.4 with a realistic endoscopic lighting condition by using a spot light attached to the main camera. We have designed a camera trajectory that hovers around the 3D models. There are a total of 900 frames of image sequences at a frame-rate of 30 fps being rendered, which is equivalent to a 30 seconds video.

Root Mean Square Distance (RMSD) is used to evaluate the overall distance between the simulated and the reconstructed surfaces. By aligning the surfaces to the real world coordinate system, we apply a grid sample to get a series of x,y coordinate points based on the surface area, and then compared the distance of the z value of the two surfaces.

The RMSD measurement for the two surface alignments has shown a good surface reconstruction results from our proposed methods, compared to the ground truth surface, the RMSD is 0.237 cm. The reconstruction error map can be viewed in Fig. 5.5.

5.3.3 Real Endoscopic Video Evaluation

To qualitatively evaluate the performance of our proposed surface reconstruction framework, we applied the proposed approach on *in vivo* videos that we acquired from Hamlyn Centre Laparoscopic / Endoscopic Video Datasets (London 2016) (Mountney and Yang 2010b), which contains 37 *in-vivo* Laparoscopic / Endoscopic videos which camera intrinsics. Fig. 5.6 (a) shows the reconstruction result from our 3D reconstruction framework with the augmented view of *in vivo* video sequences.

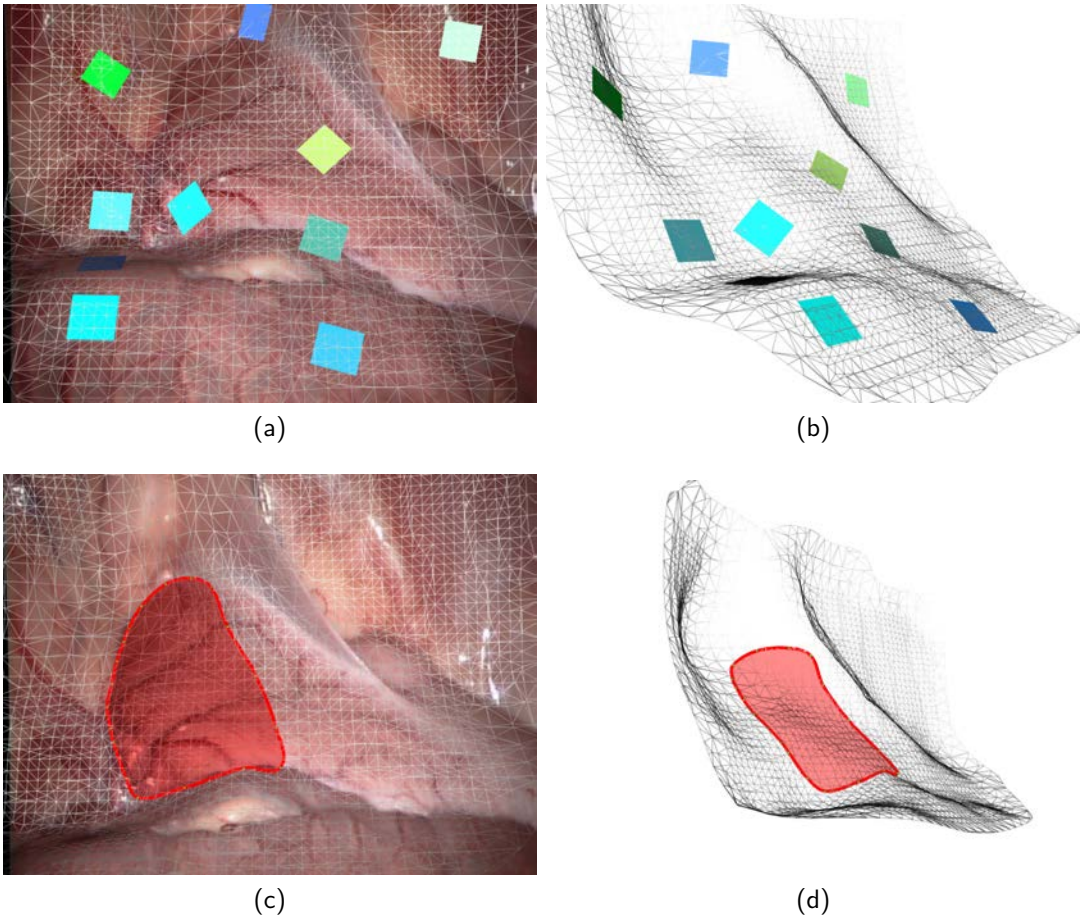


Figure 5.6: Applications of our proposed geometry-aware AR framework: (a) Adding AR labels according to the norm of the geometric surface. (b) The side-view of labels in mesh view. (c) Area highlight and measurement. (b) The side-view of highlighted area in mesh view.

By clicking the mesh, augmented objects (colored planes) can be superimposed at corresponding positions with correct poses based on the normals of the points at the click locations. Fig. 5.6 (b) shows the side-view of the mesh; note that the colored planes (which could be labels) are sticking onto the mesh correctly to create a realistic augmented environment. Fig. 5.6 (c) shows the area highlighting function of our proposed AR framework. And Fig. 5.6 (d) is the corresponding mesh view. The area highlighting function can be extended to an area measurement and line measurement (such as shown in Fig. 5.4) application once the extrinsic parameters of the camera are known.

5.4 Conclusions

In this chapter, we presented a novel AR framework for MIS. Our framework handles the two intertwined issues of tracking the rapid endoscope camera movements and providing accurate information overlay onto the curved surfaces of organs and tissues. By adapting the latest SLAM algorithms, we take a set of innovative approaches at the each stage of the AR process to improve the computational performances and AR registration accuracy. As a result, an interactive real-time geometric aware AR system has been developed. The system is capable of dealing with small soft tissue deformations, rapid endoscope movement and illumination change, which are common challenges in MIS AR. Our proposed system does not require any prior template or pre-operative scan. The system can overlay accurate augmented information such as annotations, labelling, and measurements of a tumour over curved surfaces, greatly improving the quality of AR technology in MIS.

In future work we will carry out a clinical pilot study. A case scenario will be investigated in collaboration with a practicing surgeon, and comparisons will be made as to the effectiveness of our system with the current procedural approach used.

Chapter 6

Learning-based Monocular Image Depth Estimation and 3D Reconstruction

6.1 Introduction

The human vision system is amazingly complex and extremely delicate. It can perceive depth through stereopsis, which relies on the displacement of the same object between the images received by the left and right retinas (Dunkin and Flowers 2015). With extensive visual experience and through trial and error, humans develop the ability to use contextual depth cues to achieve good and reliable perception of depth and better understanding of spatial structure. Among these depth cues, some of them do not rely on stereopsis, such as object occlusion, perspective, familiar and relative size, depth from motion, lighting and shading. Therefore, if blind in one eye or if performing a monocular task such as endoscopic surgery, we can still judge distance from these many different intuitive depth cues. In contrast, when using machine vision it is hard to infer the non-stereopsis depth cues. With the recent development of Deep Convolutional Neural Networks (DCNNs), machines can solve many computer vision problems when provided with very large human

annotated datasets such as ImageNet (Krizhevsky et al. 2012), which is known as supervised learning. Acquisition of labelled datasets is one of the biggest challenges for supervised learning, however, which is an expensive, time-consuming and labour-intensive task. In this chapter, we propose a novel self-supervised computational framework that mimics the process of how a human learns varies of contextual depth cues from stereopsis. We train a DCNN for synthesizing depth from one view of the stereo image pair, then reconstruct the other view by the synthesized depth, and finally using the stereo vision epipolar constraint (Zhang 1998) to minimize the error of the depth synthesis.

Our approach does not require the ground truth depth for supervised training. Instead, we derive the implicit function of estimating depth from monocular images by the epipolar constraint of the stereo image pair. Therefore, the method can be regarded as self-supervised learning. Compared with previous work (Garg et al. 2016) (Godard et al. 2017) (Zhou et al. 2017) addressing the same problem, we incorporate a patch-based image evaluation strategy, inspired by the classic patch matching algorithms for finding the best-matched patches between the left and right images. We use the Zero-Mean Normalized Cross Correlation (ZNCC) to measure the normalized similarities between these patches. A fully-differential patch-based ZNCC cost function is implemented to guide the depth synthesis process for more accurate results. Visual assessment shows that our approach can produce more accurate and robust depth estimations in both texture-rich and texture-less areas due to the enlargement of matching field from a pixel to a patch (see Figure 6.5). Empirical evaluations on KITTI dataset demonstrate the effectiveness of our approach

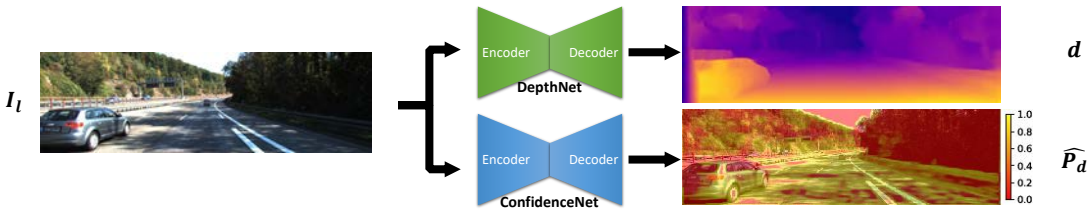


Figure 6.1: Our proposed framework can simultaneously estimate depth and the confidence of estimated depth.

and produce a state-of-the-art performance in monocular depth estimation task.

Our second contribution is that we train a parallel DCNN to evaluate the performance of the monocular depth estimation and output a 0 to 1 confidence map. The parallel DCNN is also trained in a self-supervised manner thanks to our ZNCC similarity measurement function. As ZNCC is a normalized measure of similarity, which can be approximated as the confidence of the depth estimation, we take the ZNCC loss to self-supervise the parallel DCNN (ConfidenceNet) during training so that we can estimate the confidence of the depth estimated from the first DCNN (DepthNet) during testing mode as shown in Figure 6.1. A confidence map is extremely useful for the monocular depth estimation task trained in an unsupervised manner, as the learned epipolar constraint only works well when there are clear corresponding pixels between the image pairs; it will fail and produce uncertain depth when occlusion and specularity exist in images. Our confidence map can give a basic assessment of the reliability of the predicted depth, which can then be further integrated into many applications such as monocular dense reconstruction, SLAM-based depth fusion (Tateno et al. 2017), and many tasks need crucial accurate and confidence such as monocular endoscopic surgery.

6.2 Novelty Compared to Previous Work

We propose a novel multi-scale patch-based cost function that adopts the ZNCC as a similarity function to explicitly enlarge the matching field and increase the matching robustness. From another point of view, our proposed patch-based cost function implicitly integrate the classic Patch Matching (PM) algorithm as a minimization problem in our loss function. Although Garg *et al* (Garg et al. 2016) have discussed a straightforward idea of using the stereo matching algorithm as a pre-processing method to generate "quasi ground-truth" depth for training, their result is not desirable due to the poor quality of "quasi ground-truth". Recently, Luo *et al* (Luo et al. 2018) also proposed a similar framework that firstly use a DCNN to synthesize stereo pairs from single images, and then use stereo matching to get depth. In

contrast to these two works, we treat the stereo matching as a minimization problem and implement a fully differential PM algorithm as a cost function that is seamlessly integrated into our neural network. As the loss of the PM cost function can be passed through the whole network during a backward propagation, our network can produce more robust and consistent depth by large-scale self-supervised training, which will not be limited by the performance of off-the-shelf stereo matching algorithms.

Another novelty of our work is the confidence map. As monocular depth estimation itself is an ill-posed problem, although learning-based approaches achieve comparable results to stereo depth estimation, there are still many unavoidable mistakes in the predicted depth map. For the first time, our method is able to provide a pixel-wise confidence of the predicted depth by using a parallel DCNN to capture and learn the confidence during training. The confidence map will greatly improve the usability of deploying monocular depth estimation into many practical tasks.

6.3 Method

6.3.1 Framework Overview

Figure 6.2 illustrates the entire framework for our self-supervised monocular depth learning and confidence estimation networks. Since the ground-truth depth D_{gt} is absent for supervised training, we treat the monocular depth estimation as a problem of image synthesis error minimization during training. Specifically, during training, we use the left images I_l of the stereo pairs to synthesize per-pixel depth D using an encoder-decoder network $D = F_{depth}(I_l, \theta)$, which is converted into disparities maps d by the Equation 6.2 in the next section. The disparities map d is then used to guide the stereo view reconstruction $\hat{I}_r = F_{warp}(I_l, d)$ and the sampling of patches $N_{x-d,y} = F_{sample}(I_r, d)$. After that, the loss function L_{total} is calculated based on Patch Matching Loss L_{PM}, View Reconstruction Loss L_{VR}, Disparity Smoothness Loss L_{DS}, and Disparity Consistency Loss L_{DC}. As these processes are differentiable, back propagation can be used to update the parameters θ of our depth learning

network to minimize the total loss L_{total}.

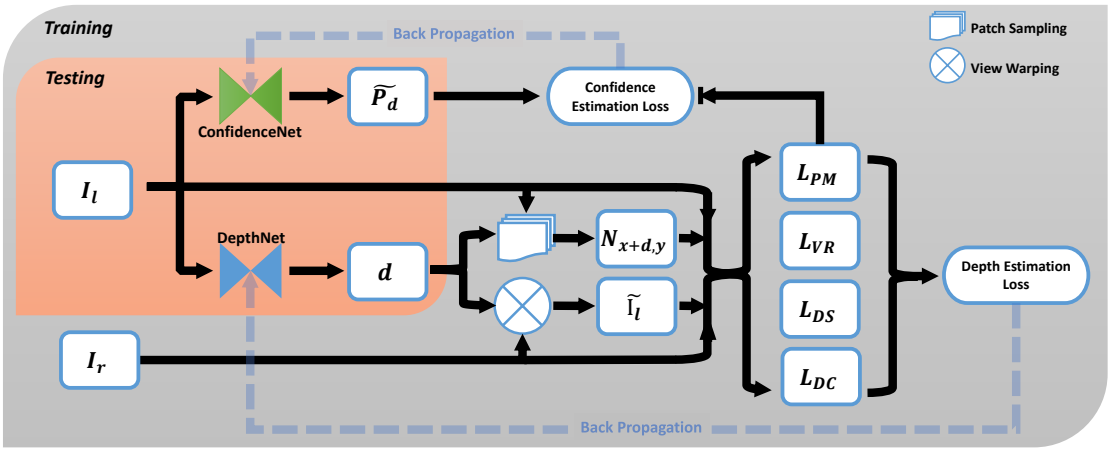


Figure 6.2: Framework for proposed self-supervised monocular depth learning and confidence estimating networks.

$$\frac{\partial L_{total}}{\partial \theta} = \frac{\partial L_{PM} + \partial L_{VR} + \partial L_{DS} + \partial L_{DC}}{\partial F_{warp}(I_l, d) + \partial F_{sample}(I_r, d)} \times \frac{\partial F_{warp}(I_l, d) + \partial F_{sample}(I_r, d)}{\partial d} \times \frac{\partial d}{\partial D} \times \frac{\partial D}{\partial \theta} \tag{6.1}$$

Since our patch-based ZNCC loss map $L_{PM}(x, y)$ represents the normalized inverted similarity between each pixel of the I_l and I_r, it can be approximated as the inverted confidence of the depth estimation result. We use the $L_{PM}(x, y)$ to self-supervise the training of a second encoder-decoder network – ConfidenceNet to generate the confidence $\hat{P}_d$ of the per-pixel depth estimation of our DepthNet.

6.3.2 Depth Synthesis Network

The core part of our framework is the depth synthesis and generation. Our goal is to learn an implicit function F_{depth} that estimates a per-pixel depth from a single input image. Inspired by the architectures of FlowNet (Dosovitskiy et al. 2017), DispNet (Mayer et al. 2016) and the network of Godard *et al* (Godard et al. 2017) and Zhou *et al* (Zhou et al. 2017), we employ a VGG-like fully convolutional neural network architecture (Shelhamer et al. 2017) in order to generate per-pixel depth

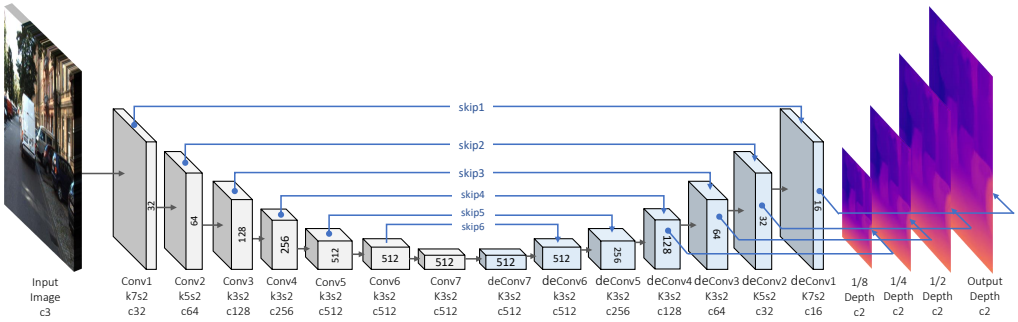


Figure 6.3: Depth synthesis network structure. "k" is the kernel size, "s" for the stride, "c" for the channel number. For simplicity, we do not draw the conv layers after each conv and deconv layer, which have the same kernel and channel size as previous layers but with stride 1.

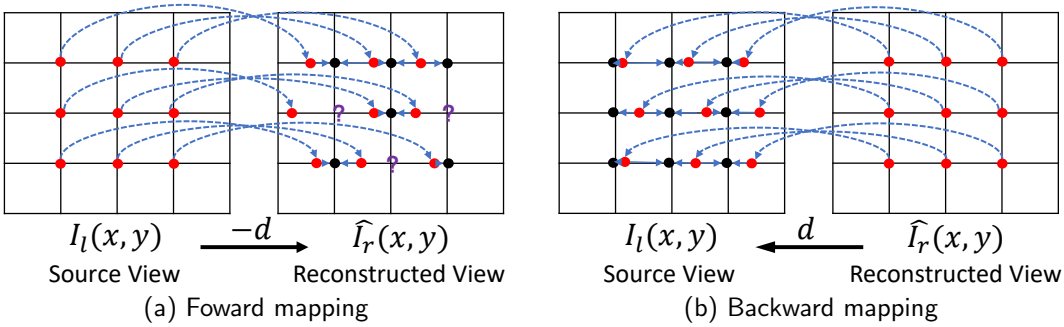


Figure 6.4: The difference between forward mapping and backward mapping.

from a single image. Our encoder-decoder model is illustrated in Figure 6.3. The input image is encoded by 7 conv layers with stride 2 each followed by a conv layer with stride 1, which efficiently compress the input image into a feature tensor with $1/2^7$ original size and 512 channels. Then, the feature tensor is up-sampled by 7 deConv layers with stride 2 each followed by a conv layer with stride 1, which decode the feature tensor into a full original size depth. Following the method in (Dosovitskiy et al. 2017), 6 skip connections are implemented for preserving high-level information to ensure the high quality per-pixel prediction after up-sampling. Multi-scale depth images are outputted and used for further steps to constraint the network for a coarse-to-fine up-sampling.

6.3.3 Warping-based Stereo View Reconstruction

View warping is an enabling technology for self-supervised learning framework (Garg et al. 2016) (Godard et al. 2017) (Zhou et al. 2017). Given the per-pixel disparity map estimated from a single image in the previous step (calculated by Equation 6.2), the target view of the stereo pairs can be reconstructed by the epipolar relationship in stereo vision. According to the epipolar constraint: the projection of a pixel x_l on the right camera plane x_r must be contained in the epipolar line. For calibrated stereo pairs discussed in this chapter, x_l and x_r must be in the same row y, and the disparity d describes the horizontal displacement of the corresponding pixels x_l and x_r . Through the stereo triangulation, we can get that

$$D_{xy} = \frac{bf}{d} \;\Rightarrow\; d = x_l - x_r = \frac{bf}{D_{xy}} \tag{6.2}$$

where D_{xy} is the depth estimated in the pixel at (x, y), b and f are the camera baseline and focal distance. By the relationship discussed in the above equation, the target view in a stereo pair can be reconstructed given the source view and the corresponding depth (estimated through our depth synthesis network).

However, the direct mapping from one known view to the other view (forward mapping) will result in holes in the target image that are not differentiable (see Figure 6.4 (a)). Therefore, we use the inverse mapping (Figure 6.4 (b)): for each pixel in the target view, by picking points from the source to reconstruct the target view guided by the d. Thus, a relatively more complete and differentiable target view can be generated than the foward mapping. Then the bilinear sampling (Jaderberg et al. 2015) is used to get the interpolated pixel value from the source view.

6.3.4 Disparity-guided Patch Sampling

Inspired by the stereo view reconstruction described above, we propose a novel patch sampling process guided by the estimated disparity from our DepthNet. $N_{x,y}$ is defined as a patch with window size n, centered at the coordinate (x, y). We sample patches on each pixel in the left image $\{x, y \in I_l | I_l(N_{x,y})\}$, and the cor-

responding patches shifted by disparity values d of each pixel in the right image, $\{x, y \in I_r | I_r(N_{x-d,y})\}$. According to Equation 6.2, if d is correct, then we have $I_l(N_{x,y}) = I_r(N_{x-d,y})$. And this relationship will be used to construct the patch matching loss. These sampled patches are computed and stored vectorized so that can be deployed parallelly on GPU for accelerated computation.

The patch sampling size is very important and can affect the final performance of similarity measurement. However, there is no optimal patch size and the performance varies greatly across different images and local details. When small patch size is used, little information will be captured, and the similarity comparison robustness will be decreased. If we use a large patch size, computational complexity will be greatly increased and also cannot recover accurate depth at stereo occlusion and depth discontinuous. Therefore, we use a multi-scale patch sampling scheme and sample a combination of 4 different patch sizes in an image to fully exploit the effects of different patch sizes. We will discuss the choice of patch sizes in Section 6.4.1.

6.3.5 Loss Function Construction

We define a loss function L_{total} with multiple strategies to effectively train our networks for accurate, smooth and realistic depth.

$$L_{total} = \omega_p L_{PM} + \omega_v L_{VR} + \omega_d L_{DS} + \omega_c L_{DC} \tag{6.3}$$

where from left to right is: Patch Matching Loss, View Reconstruction Loss, Disparity Smoothness Loss and Disparity Consistency Loss. $\omega_p, \omega_v, \omega_d, \omega_c$ are the corresponding weights for the Patch Matching Loss, View Reconstruction Loss, Disparity Smoothness Loss and Disparity Consistency Loss, to balance the effects of gradients back propagation. Each loss function will be explained in details below:

6.3.5.1 Patch Matching Loss

Inspired by patch matching algorithm that by finding the best-matched patches in the left and right image to get correct disparities. We propose a patch matching loss that can be used to maximize the similarities (minimize the differences) of patches in left image $I_l(N_{x,y})$ and the shifted patches in right image $I_r(N_{x-d,y})$ to get correct disparities. Here, the ZNCC measure of similarity is used to compute a normalized similarity between the patches $I_l(N_{x,y})$ and $I_r(N_{x-d,y})$:

$$C_{ZNCC}\left(I_l(N_{x,y}), I_r(N_{x-d,y})\right) = \frac{\sum_{i,j\in N_{x,y}}\left(I_l(i,j)-\bar{I}_l(N_{x,y})\right)\cdot\left(I_r(i-d,j)-\bar{I}_r\left(N_{x-d,y}\right)\right)}{\sqrt{\sum_{i,j\in N_{x,y}}\left(I_l(i,j)-\bar{I}_l(N_{x,y})\right)^2\cdot\sum_{i,j\in N_{x,y}}\left(I_r(i-d,j)-\bar{I}_r\left(N_{x-d,y}\right)\right)^2}} \quad (6.4)$$

where $\bar{I}\left(N_{x,y}\right) = \frac{1}{n}\sum_{x,y\in N_{x,y}} I\left(x,y\right)$ is the mean intensity of the patch $N_{x,y}$ centered at the coordinate (x,y).

The ZNCC returns a similarity ranging from $[-1,1]$. We first normalize it into $[0,1]$ then invert it to get the patch matching loss:

$$L_{PM} = \sum_{x,y} 1 - \frac{1 + C_{ZNCC}\left(I_l(N_{x,y}), I_r(N_{x-d,y})\right)}{2} \quad (6.5)$$

Our patch matching loss is computed at all 4 patch sizes to cover both small structures and large areas. There are several advantages of using our patch-based ZNCC loss to regularize the depth synthesis:

(1) Our patch matching loss uses patches for measurement that involve larger regions than the direct pixel-wise photometric loss used in previous work, which is more robust and can achieve sub-pixel accuracy. Figure 6.5 demonstrates the effect of our patch-based ZNCC loss. We charted the values of our patch-based ZNCC loss and the photometric loss against the disparity value of a pixel located at the center of the image patch "6". It is obvious that by using our proposed patch-based ZNCC loss, the loss is more smooth and likely to converge to the global minimum. Whereas the direct pixel-wise photometric loss will lead to many local minimums shown in the right curve chart.

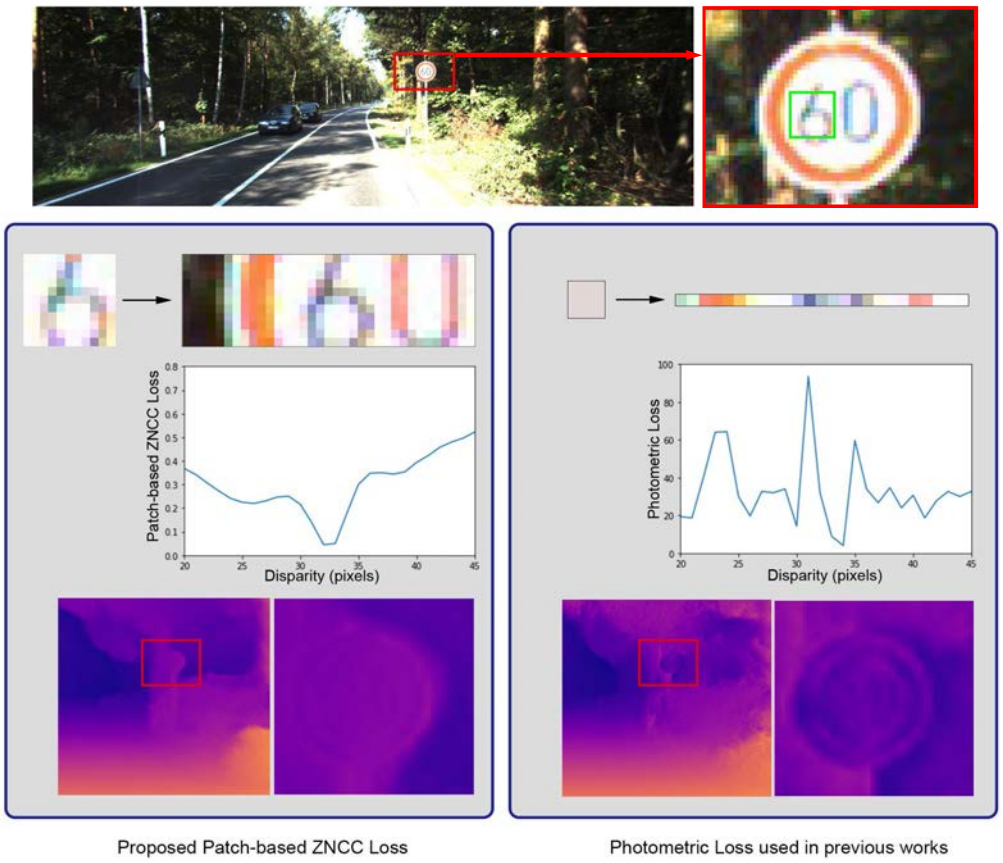


Figure 6.5: Comparison of our proposed patch-based ZNCC loss with the photometric loss used in previous works.

(2) Compared to other similarity measures such as absolute intensity difference (AD), Census, and Normalized Cross Correlation (NCC), ZNCC is especially robust against Gaussian noise and variation between the compared patches, which can help to recover more accurate depth in our self-supervised framework.

(3) As a zero-mean normalized similarity measurement function, our patch-based ZNCC loss can provide a similar value ranging from $[-1, 1]$. After normalized to $[0, 1]$ as shown in Equation 6.5, it can be regarded as the confidence of the generated depth at each pixel, which can be further used to self-supervise the training of our confidence network.

6.3.5.2 View Reconstruction Loss

We use the view reconstruction loss as a second supervision on the depth synthesis. Guided by the synthesized depth, the right views can be reconstructed by collecting pixels from left images. The view reconstruction loss is defined as the L1 loss between the reconstructed view $\hat{I}_r$ and the original view I_r:

$$L_{VR} = \sum_{xy} \left| I_r(x,y) - \hat{I}_r(x,y) \right| \tag{6.6}$$

Compared to the patch matching loss, the view reconstruction L1 loss is more sensitive to small structures and depth discontinuities and can provide more detailed depth information.

6.3.5.3 Disparity Smoothness Loss

We use a disparity smoothness term to regularize our network to produce more smooth depth. Similar to (Garg et al. 2016) (Godard et al. 2017) (Zhou et al. 2017), we use the sum of the L1 norm of the disparity gradients along the x and y directions as a smoothness factor. The edge-aware terms are used to reduce the penalty on edges where depth discontinuities usually happen, which can prevent over-smoothing.

$$L_{DS} = \frac{1}{XY} \sum_{x,y} \left| \frac{\partial d(x,y)}{\partial x} \right| e^{-\| \frac{\partial I(x,y)}{\partial x} \|} + \left| \frac{\partial d(x,y)}{\partial y} \right| e^{-\| \frac{\partial I(x,y)}{\partial y} \|} \tag{6.7}$$

6.3.5.4 Disparity Consistency Loss

The left-right disparity consistency loss proposed in (Godard et al. 2017) has achieved a great improvement for monocular depth generation. Here, we adopt this loss function into our framework. The left and right image disparities are both generated, and the difference of left disparity map and the reconstructed left disparity map from right disparity is computed and minimized. This loss will ensure the left and

right disparities coherence.

$$L_{DC} = \frac{1}{XY} \sum_{x,y} |d_l(x,y) - d_r(x - d_l(x,y), y)| \tag{6.8}$$

6.3.6 Confidence Estimation Network

One of the advantages of our proposed patch matching loss is that a normalize similarity measurement can be generated for each pixel at the training time. With the well-known epipolar constraint, the per-pixel confidence of the estimated depth can be approximated as the normalized similarity measurement of the left patches and the corresponding patches in the right image.

$$P_d(x,y) \approx C_{Normalized}(I_l(N_{x,y}), I_r(N_{x-d,y})) = (1 - L_{PM}(x,y)) \tag{6.9}$$

Here, we propose to use another encoder-decoder network to learn the confidence map generated by our depth estimation network during training, so that the confidence map can be preserved and generated during the testing time. We tried to train the confidence and depth in one network like (Ladický et al. 2014) (Eigen and Fergus 2015) (Wang et al. 2015c) (Mousavian et al. 2016), but the multi-task training would reduce the depth estimation performance. Therefore, we use a parallel encoder-decoder network to learn the confidence supervised by the per-pixel ZNCC loss of our depth estimation network. The loss of our ConfidencNet is shown below:

$$L_{ConfidenceNet} = \sum_{x,y} \left| (1 - L_{PM}(x,y)) - \hat{P}_d(x,y) \right| \tag{6.10}$$

where $\hat{P}_d(x,y)$ is the generated confidence map, $L_{PM}(x,y)$ is the patch matching loss from our depth estimation network described in above sections. The static copy is used here to prevent the gradients propagating back to the depth estimation network. The $1 - L_{PM}(x,y)$ operation inverts the loss to confidence, and L1 loss is used to access the confidence estimation error.

Instead of using the same encoder-decoder network structure as our DepthNet,

we employ a simpler structure by only using first 5 conv-layer and last 5 deconv-layer without skip layers as described in Figure 6.3 for two reasons:

(1) To reduce memory usage and training time, as training two neural networks at the same time is very computationally expensive. The second network can be replaced by a deeper and more complex encoder-decoder network to produce sharper and more accurate confidence, but the main purpose of our work is to prove that our self-supervised monocular depth learning and confidence estimation framework is feasible and helpful for depth prediction, hence we choose to use a simple network structure as the proof of concept.

(2) We intend to use a simpler network with fewer weights to prevent over-fitting to noises and to learn more generic confidence – high confidence in texture-rich areas, low confidence in texture-less, blurry and occluded areas, which is what we design this confidence net for.

6.4 Experiments

In this section, we evaluate our framework and compare the results with prior approaches both quantitatively and qualitatively on KITTI dataset. We use the rectified stereo image pairs for training our networks. For testing time, we use the left image to generate depth, and the corresponding sparse LIDAR data is served as the ground truth for benchmarking.

6.4.1 Implementation Details

Our networks are implemented in Tensorflow and trained on a workstation with a single Nvidia Titan X GPU (12G Memory). Our models take around 60 hours to train for 50 epochs. When in testing mode, our networks can output depth and confidence map at around 20 frames per second. We open-sourced our implementation which can be downloaded from here [1].

[1] `https://github.com/melights/Monodepth_with_Confidence`

Hyper Parameters. All input images are scaled to 512x256 with a batch size of 4. Adam Optimizer is used with $\beta_1 = 0.9$, $\beta_1 = 0.999$, and initial learning rate $\lambda = 0.0001$ that decays after half of the training process. The weights to construct our total loss function for depth estimation network are $w_p = 0.5$,$w_v = 1$,$w_d = 0.1$,$w_c = 1$.

Data Augmentation. The same data augmentation approach in (Godard et al. 2017) is used to randomly flip the image and change the gamma, brightness, and color shifts to increase the network robustness and prevent over-fitting.

Multi-scale Implementation. We employ a multi-scale strategy to ensure a coarse-to-fine up-sampling. As can be seen from Figure 6.3, 4 depth scales are outputted with $1/8, 1/4, 1/2$ and a full resolution. All of our loss functions are computed for each of these 4 scales, and for each of left and right images/disparities. We take the means of these loss functions as the final loss.

Patch Size. By applying different patch sizes on different image scales, we can get very large equivalent patch sizes with less computation. For patch size choices, based on our empirical test, we use $n = 5, 5, 7, 9$ pixels for our patch-based ZNCC loss on 4 different scales, which is equivalent $n = 5, 10, 28, 72$ pixels' windows on full resolution images.

6.4.2 KITTI dataset

To be able to compare with the state-of-the-art monocular depth learning approaches, we trained and evaluated our networks using two different train/test splits: *Godard* and *Eigen*.

Godard Split. We use the same train/test sets that Godard *et al* (Godard et al. 2017) proposed in their work. 200 high quality disparity images in 28 scenes provided by the official KITTI training set are served as the ground truth for benchmarking. For the rest of 33 scenes with a total of 30,159 images, 29,000 images are picked for training and the remaining 1,159 images for testing.

Eigen Split. For fair comparison with more previous works, we also use the test split proposed by Eigen *et al* (Eigen et al. 2014) that has been widely evaluated

Table 6.1: Comparison with state-of-the-art methods on KITTI dataset.

Method	Super-vision	Split	Cap	Error (Lower better)					Accuracy (Higher better)		
				AbsRel	SqRel	RMSE	RMSElog	D1-all	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen et al (Eigen et al. 2014)	Yes	E	80	0.203	1.548	6.307	0.282	-	0.702	0.890	0.958
Liu et al (Liu et al. 2016)	Yes	E	80	0.201	1.584	6.471	0.273	-	0.680	0.898	0.967
Zhou et al (Zhou et al. 2017)	No	E	80	0.208	1.768	6.856	0.283	-	0.678	0.885	0.957
Godard et al (Godard et al. 2017)	No	E	80	0.148	1.344	5.927	0.247	-	0.803	0.922	0.964
Ours	**No**	**E**	**80**	**0.145**	**1.267**	**5.786**	**0.244**	**-**	**0.811**	**0.925**	**0.965**
Garg et al (Garg et al. 2016)	No	E	50	0.169	1.080	5.104	0.273	-	0.740	0.904	0.962
Zhou et al (Zhou et al. 2017)	No	E	50	0.201	1.391	5.181	0.264	-	0.696	0.900	0.966
Godard et al (Godard et al. 2017)	No	E	50	0.140	0.976	4.471	0.232	-	0.818	0.931	0.969
Ours	**No**	**E**	**50**	**0.138**	**0.937**	**4.399**	**0.231**	**-**	**0.825**	**0.933**	**0.969**
Godard et al (Godard et al. 2017)	No	G	80	0.124	1.388	6.125	0.217	30.272	0.841	0.936	0.975
Ours	**No**	**G**	**80**	**0.117**	**1.202**	**5.953**	**0.210**	**29.612**	**0.845**	**0.938**	**0.976**

by the works of Garg *et al* (Garg et al. 2016), Liu *et al* (Liu et al. 2016), Zhou *et al* (Zhou et al. 2017) and Godard *et al* (Godard et al. 2017). This test split contains 697 images of 29 scenes. The rest of 32 scenes contain 23,488 images, in which 22,600 are used for training and the remaining for testing, similar to (Garg et al. 2016) and (Godard et al. 2017).

6.4.3 Results

6.4.3.1 Quantitative Evaluation

The evaluation results on the KITTI dataset are reported in Table 6.1. We use different combinations of train/test splits (E for Eigen, G for Godard) and cap distances (80m and 50m) to compare with different works. For Eigen *et al* (Eigen et al. 2014), Liu *et al* (Liu et al. 2016), Zhou *et al* (Zhou et al. 2017) and Godard *et al* (Godard et al. 2017) , the Eigen split with 80m cap distance are used. For Garg *et al* (Garg et al. 2016), Zhou *et al* (Zhou et al. 2017) and Godard *et al* (Godard et al. 2017), the Eigen split with 50m cap distance are used. We also report our result on Godard split with 80m cap. The results show that our method outperforms all compared methods and produce the state-of-the-art results for monocular depth estimation problem on KITTI dataset in terms of error and accuracy metrics.

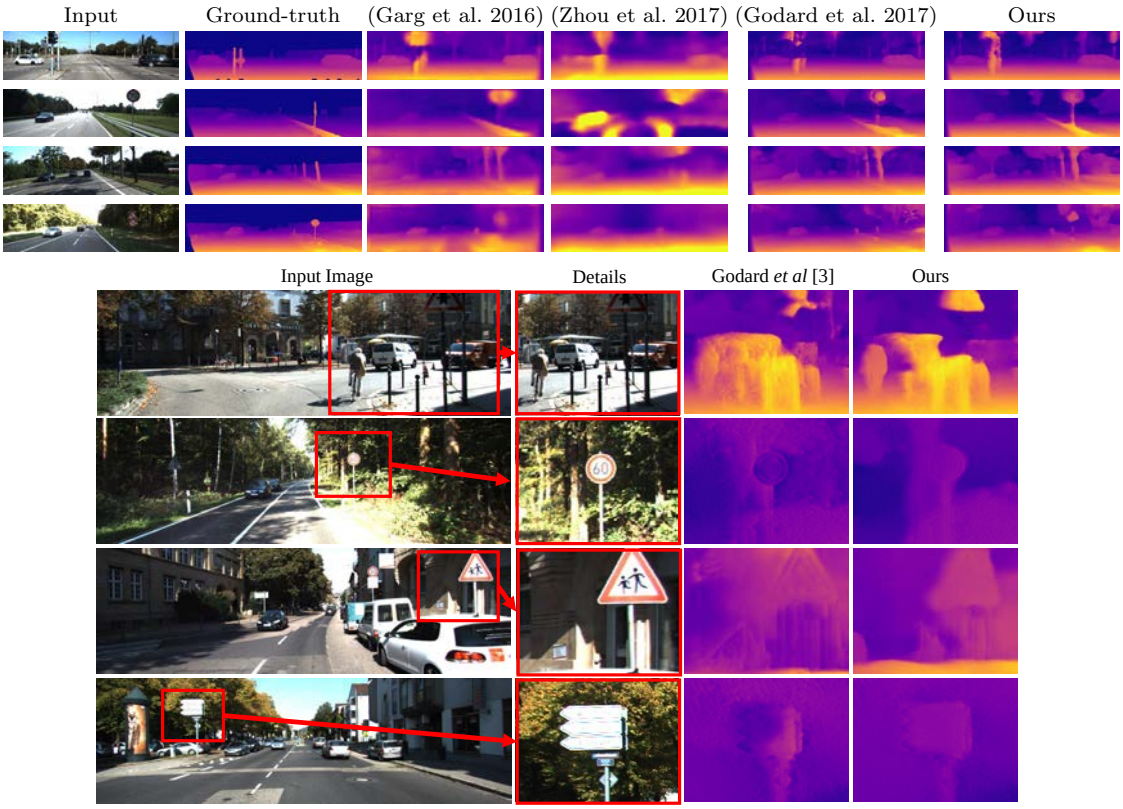


Figure 6.6: Upper part: comparison of monocular depth estimation on KITTI dataset between Garg *et al*(Garg et al. 2016), Zhou *et al*(Zhou et al. 2017), Godard *et al*(Godard et al. 2017), and ours. Lower part: comparison of details with Godard *et al*(Godard et al. 2017). All of the results are generated using authors' provided pre-trainned model. The ground-truth depth map is interpolated from sparse point map only for visualization.

6.4.3.2 Qualitative Evaluation

The qualitative comparison to some of the related methods on KITTI dataset is shown in Figure 6.6. While our network structure is similar to that of Godard *et al*(Godard et al. 2017), both generate clear and accurate depth than other works. We also provide a detailed comparison with the results of Godard *et al*(Godard et al. 2017) in the lower part of Figure 6.6. Our network can generate more accurate depth in complex regions with thin structures and texture-less areas such as the pillars and traffic signs. This verified the theory we explained in Figure 6.5 that our patch-based loss function is more robust and easier to converge to the global minimum in complex regions.

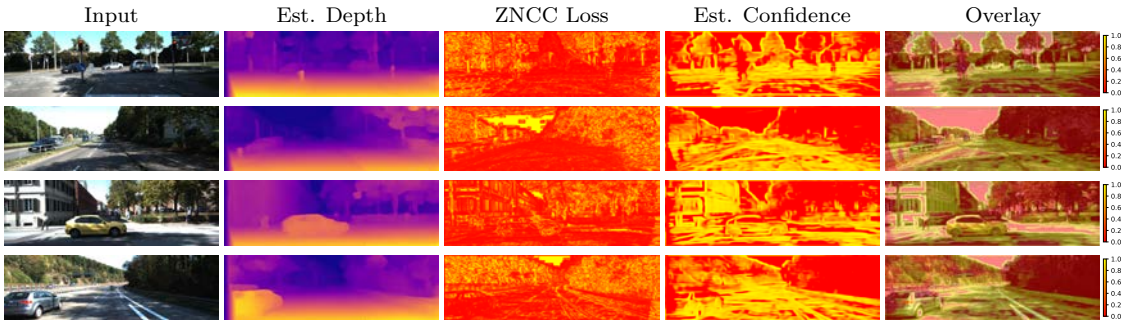


Figure 6.7: Confidence estimation results. A colorbar from red to yellow is used to represent 0 to 1.

6.4.3.3 Confidence Map Evaluation

We show the confidence estimation results in Figure 6.7. A colorbar from red to yellow is used to represent 0 to 1. We can see that the estimated confidence can nicely represent the inverted ZNCC loss but less noisy due to the small network we use to prevent over-fitting. The overlaid confidence on input image shows that our ConfidenceNet has learned to generate confidence from contextual information. For example, in texture-less areas (sky, building), dark areas (trees under shadow), occluded areas (around thin structures) and reflective areas (car window), the estimated confidence is usually very low, while the texture-rich areas and edges usually have high confidence.

6.4.3.4 Reconstruction Results

As can be seen in Figure 6.8, we compared the reconstruction results based on the depth map predicted using Stereo Matching (Chen et al. 2001), Godard *et al*(Godard et al. 2017), and ours proposed method. Stereo Matching cannot estimate correct depth for some pixel and left many holes which results in bad reconstruction result. For the reconstruction results of Godard *et al*(Godard et al. 2017), as discussed before, cannot recovery correct depth in the areas that contain small structures, which results in noisy reconstruction result. Our method can generate correct and smooth depth and reconstruction.

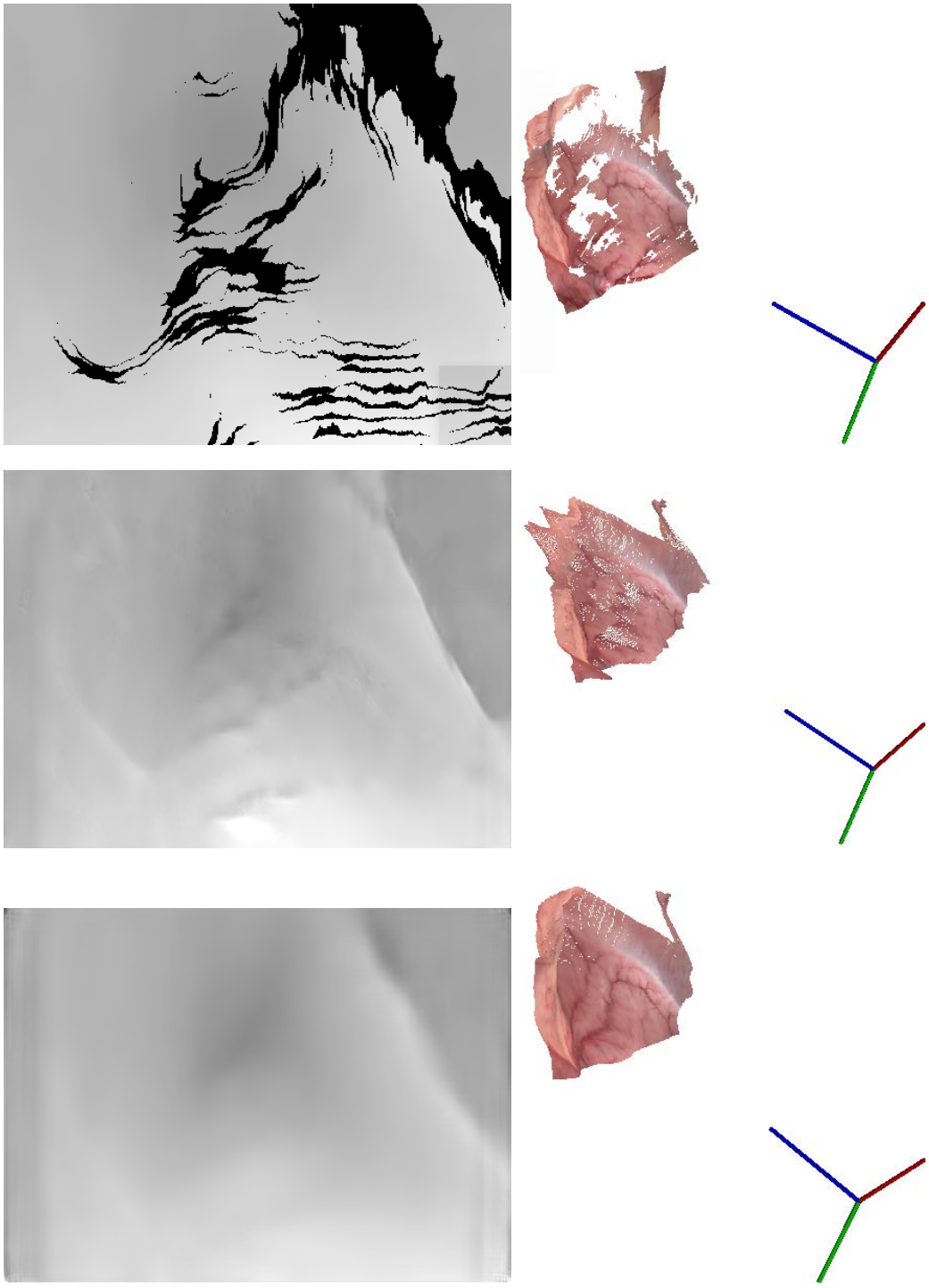


Figure 6.8: Reconstruction result (right) based on the estimated depth (left) from different approaches (from top to bottom: Stereo Matching (Chen et al. 2001), Godard *et al*(Godard et al. 2017), and ours.)

6.5 Discussion

In this chapter, we have presented a novel self-supervised framework for monocular depth learning and confidence estimation. We incorporate the patch matching theory into a fully differential DCNN and achieve self-supervised training of both depth and the confidence of depth. Our proposed loss function exploits the epipolar constraint of stereo vision and also provides a normalized similarity that is further used to supervise the confidence estimation. Our method not only outperforms the state-of-the-art results on the KITTI benchmark evaluation, but also for the first time, we are able to simultaneously generate depth from monocular images and estimate the confidence of the generated depth. This is a step change for monocular depth estimation as it significantly increases the feasibility of using monocular depth estimation into many practical applications such as autonomous driving and monocular endoscopic surgery, where the accuracy of estimated depth is crucial.

Why Our ConfidenceNet Works? Since our ConfidenceNet is supervised by the per-pixel ZNCC loss of our depth estimation network, it explicitly learns the regions where our depth estimation network performs well and badly. But on a deeper level, our ConfidenceNet actually implicitly learns the inherent defect of the patch matching algorithm – it would fail on texture-less regions and performs badly near stereo view occlusions, reflections and blurred areas. Therefore, after sufficient training steps, our ConfidenceNet can give an estimation of the confidence of our DepthNet, although they are two different networks.

In Future Work We will continue optimizing our model and explore the possibility of using adaptive window size for patch sampling to decrease the training time and increase accuracy in small structures.

Chapter 7

From Geometry-Aware AR to Context-Aware AR

7.1 Introduction

Mixed Reality combines computer vision with information science, and computer graphics as a cross-cutting technology. It makes seamless connections between virtual space and the real world, by not only superimposing computer-generated information onto the real world environment, but also making progress on novel user interaction for new experience. This interactive technology will soon become ubiquitous in many applications, ranging from personal information systems, industrial and military simulations, office use, digital games to education and training.

The latest research on Simultaneous Localisation and Mapping (SLAM) has opened up a new world for MR development, greatly increased the camera tracking accuracy and robustness. The sparse SLAM systems (Davison et al. 2007) (Klein and Murray 2007) (Mur-Artal et al. 2015) are proven to be efficient 3D tracking methods for monocular cameras, but structural information are absent from these systems. In contrast, dense SLAMs (Newcombe et al. 2011b) (Newcombe et al. 2011a) (Newcombe et al. 2015) construct dense surfaces to generate geometric information of the real scene, enabling geometric interactions in MR environment. The collision effects between virtual and real-world objects in these geometry-aware MR

systems do increase the immersion of the user experience (as can be seen in Figure 7.1 (a) and (b) for the Ball Pit MR game in Microsoft HoloLens). However, as individual semantic properties of various different objects of the real world remain undetected, geometry-aware MRs are unable to distinguish different object properties and may always generate uniform interactions between each other, which will break the continuous user experience in MR (Grubert et al. 2017).

The natural first step moving away from purely geometric-based approaches towards generating context-aware interactions is to understand the real environment semantically in MR. Semantic segmentation (Garcia-Garcia et al. 2017) (Shelhamer et al. 2017) (Zheng et al. 2015) (Chen et al. 2017c) (Badrinarayanan et al. 2017) leading to semantic understanding is not new to computer vision. However, very few prior works of utilizing semantic information in MR are reported. Semantic understanding in MR presents additional challenges (1) associating semantics with the structural information of the environment seamlessly on-the-go and (2) retrieving the semantics then generating appropriate interactions.

Embedding semantic information extracted from a 2D image space into the 3D structure of a MR environment is hard, because of the required high accuracy. Careful considerations are needed when designing semantic-based MR interactions.

Realistic interactions in MR require not only geometric and structural information, but also semantic understandings of the scene. While geometric structure allows accurate information augmentation and placement, at the user experience level, semantic knowledge will enable realistic interactions between the virtual and real objects. For example realistic physical interactions (e.g. a virtual glass can be shunted on a real concrete floor) in MR. More importantly, using semantic scene descriptions, we can develop high-level tools for efficient design and constructions of large and complex MR applications.

With the gap that this work addresses, we propose a novel context-aware semantic MR framework and demonstrate its effectiveness through example interactive applications. We show that how an end-to-end Deep Learning (DL) framework and the dense Simultaneous Localisation and Mapping (SLAM) can be used for semantic

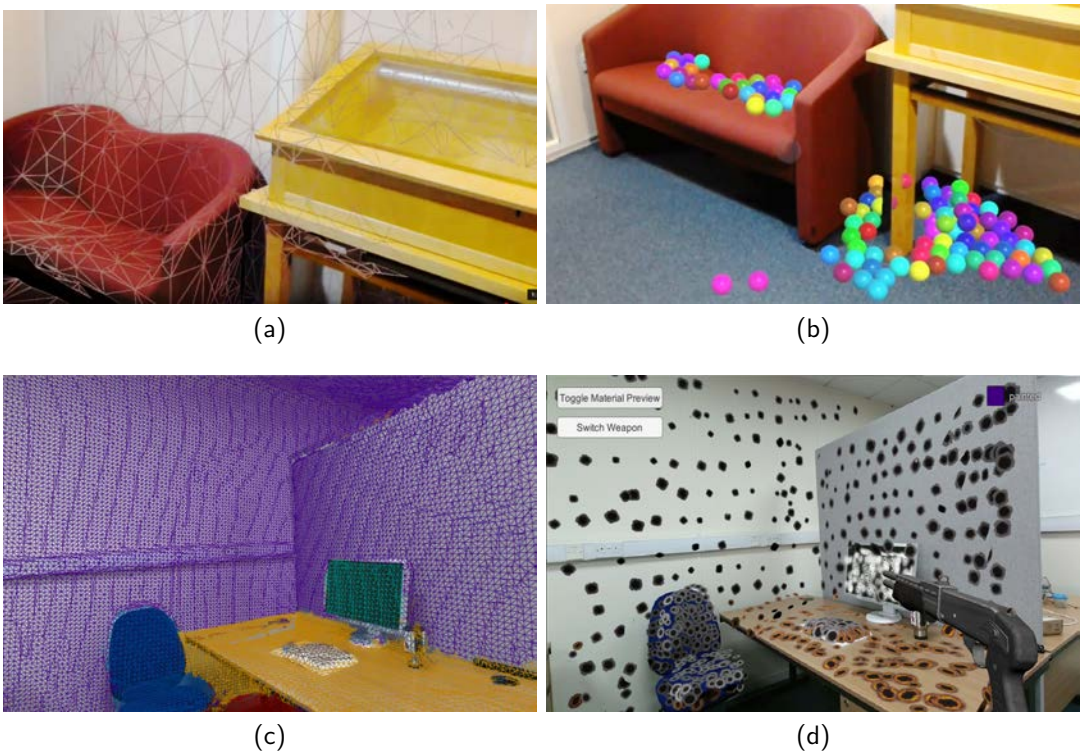


Figure 7.1: (a): Microsoft HoloLens is capable of reconstructing the environment by its built-in "spatial mapping" function and provide a geometric mesh for geometry based interaction. (b): The Ball Pit MR games based on geometry interaction for Microsoft HoloLens. (c) Our proposed framework can provide semantic mesh for more advanced context-aware interaction. (d) A shooting game developed based on our proposed framework. Note that the bullet holes are different according to the objects' properties.

information integration in MR environment and how context-aware interactions can be generated. We present the labelling of material properties of the real environment in 3D space as a novel example application to deliver realistic physical interactions between the virtual and real objects in MR. Our key insight is to build semantic understanding in MR that not only can greatly enhance user experience through object-specific behaviours, but also pave the way for solving complex interaction design challenges. To the best of our knowledge, this is the first work to present a context-aware MR (1) using deep learning based semantic scene understanding and (2) generating semantic interactions at the object-specific level, one step further towards the high-level conceptual interaction modelling in complex MR environment

for enhanced user experience.

Dense SLAM KinectFusion (Newcombe et al. 2011a) is used for camera pose recovery and 3D model reconstruction for creating a classic geometry-aware MR environment first. We trained a Conditional Random Fields as Recurrent Neural Networks (CRF-RNN) (Zheng et al. 2015) using a large-scale material database (Bell et al. 2015) for detecting material properties of each object in the scene. The 3D geometry/model of the scene is then labelled with the semantic information about the materials made up of the scene, so that realistic physics can be applied during interactions via real-time inference to generate corresponding physical interactions based on the material property of the object that the user is interacting with, as shown in a shooting game example for object-specific material-aware interactions, Figures 7.1 (c) and (d).

The framework is both efficient and accurate in semantic labelling and inference for generating realistic context-aware interactions. Two tests are designed to evaluate the effectiveness of the framework (1) an accuracy study with an end-to-end system accuracy evaluation by comparing the dense semantic ray-casting results with manually labelled ground truth from 25 keyframes of two different scenes and (2) a user experiment with 68 participants to qualitatively evaluate user experience using three different MR conditions. The results show that the framework delivers more accuracy in 3D semantic mappings than directly using the state-of-the-art 2D semantic segmentation. The proposed semantic based interactive MR system ($M = 8.427, SD = 1.995$) has a significant improvement ($p < 0.001$) on the realism and user experience than the existing MR system approaches that do not encode semantic descriptions and context-aware interaction modelling ($M = 5.935, SD = 1.458$).

In the next section, we review related work on geometry-based MR Interactions, and recent approaches to semantic segmentations using Convolutional Neural Network. The following sections introduce our framework with SLAM dense reconstructions of the scene and the 3D semantic fusion, and describe our implementation and evaluation framework. Finally, we demonstrate our results compared with the

state-of-the-art semantic segmentation algorithms.

7.2 Previous Work

Our approach draws on recent success of dense SLAM algorithms (Newcombe et al. 2011b) (Newcombe et al. 2011a) (Newcombe et al. 2015) and deep learning for semantic segmentations (Garcia-Garcia et al. 2017) (Shelhamer et al. 2017) (Zheng et al. 2015) (Chen et al. 2017c) (Badrinarayanan et al. 2017) that have been mostly used in the field of robotics until now.

7.2.1 Geometry-based MR Interaction

Interaction modelling between virtual and real objects in MR are mostly geometry-based through plane feature detections or full 3D reconstructions of the real-world. Methods of using plane detections (Salas-Moreno et al. 2014) (Nuernberger et al. 2016) estimate planar surfaces in the real-world, onto which virtual objects are placed and collided with. Random Sample Consensus (RANSAC) algorithm (Fischler and Bolles 1981b) estimates planar surfaces based on sparse 3D feature points extracted from a monocular camera. Plan detections require no depth cameras, are computationally efficient and run on mobile phones. Mobile MR experience is shown in the newly released Mobile AR systems (Apple 2017) (Google 2017). One obvious shortcoming of plane detections is the requirement for large planar surfaces to delivery MR interactions. Collision meshes for non-planar surfaces are impossible, hence, restricting user experience to the area and types of objects which users can interact with.

Recent advances in depth sensors, display technologies and SLAM software (Newcombe et al. 2011b) (Newcombe et al. 2011a) (Whelan et al. 2013) (Newcombe et al. 2015) have opened up the potential of MR systems. Spatial structures of the real environment can be generated at ease to provide accurate geometries for detecting collisions between virtual and real objects. We saw examples of geometry-based interactions e.g. a virtual car 'drives' on an uneven real desk (Newcombe et al.

2011b); the Super Mario game played on real building blocks (Kim et al. 2013); and the Ball Pit game in HoloLens (Microsoft 2017). Figures 7.1 (a) and (b) illustrate the concept. Impressive as they are, the state of the art systems are still limited to the basic and uniform geometry-based virtual and real object interactions. Without high-level semantic descriptions and scene understandings, continuous user experience in MR is compromised and easily broken, and the realism and immersion are reduced. One example is in the Ball Pit game, material properties of the real objects are not recognized, thus a ball falling onto a soft surface would still bounce back unrealistically against the law of physics.

7.2.2 Deep Semantic Understanding

Semantic segmentation is an emerging technology in computer vision. The recent success of Convolutional Neural Network(CNN) has achieved the semantic level image recognition and classification with great accuracy (Krizhevsky et al. 2012), enabling many novel applications. In last few years more complex neural networks such as FCN (Shelhamer et al. 2017), CRF-RNN (Zheng et al. 2015), DeepLab (Chen et al. 2017c) and SegNet (Badrinarayanan et al. 2017) have enabled image understanding at the pixel level. Semantic information at every pixel of an image can be predicted and labelled when using these neural networks trained on large-scale databases.

Combined with SLAM systems, semantic segmentation can be achieved in 3D environments (Rünz and Agapito 2017) (Tateno et al. 2017) (Zhao et al. 2017) (McCormac et al. 2017), a promising future in robotic vision understanding and autonomous driving. Unlike these existing methods that aimed at providing semantic understanding of the scene for robots, we focus on human interactions. Our goal is to provide users with realistic semantic level interactions in MR. In this chapter, for the first time, we use MR as a bridge to connect AI and human for a better understanding of the world via intelligent context-aware interaction.

7.2.3 Context and Semantic awareness in XR environment

Prior approaches have studied context and semantic understandings in 3D virtual environment, e.g. semantic inferring in interactive visual data exploration (North et al. 2012); enhancing software quality for multi-modal Virtual Reality (VR) systems (Fischbach et al. 2017); visual text analytics (Endert et al. 2012); and interactive urban visualization (Deng et al. 2016). Context awareness is also introduced in computer-aided graphic design such as inbetweening of animation(Yang 2018); 3D particle clouds selection(Yu et al. 2016); and illustrative volume rendering (Rautek et al. 2007). Virtual object classifications are proposed in VR applications by using semantic associations to describe virtual object behaviours (Chevaillier et al. 2012). The notion of *conceptual modelling* for VR applications is pointed out by Troyer *et al.*, highlighting a large gap between the conceptual modelling and VR implementations. It is suggested to take a phased approach (i.e. conceptual specification, mapping and generation phases) to bridge the gap (De Troyer et al. 2007).

Recently, the idea of extending Augmented Reality (AR) applications to become context-aware has appeared in computer graphics (Grubert et al. 2017), which proposes to classify context sources and context targets for continuous user experience. A method is proposed for authentically simulating outdoor shadows to achieve seamless context-aware integration between virtual and real objects for mobile AR (Barreira et al. 2018).

We address ubiquitous interactions in MR environment and see deep semantic understanding of the environment as the first step towards the high-level interaction design for MR. Real-time 3D semantic reconstruction is an active research topic in robotics with many recent works being focused on object semantic labelling. We now re-design and fine-tune the architecture for this MR framework.

7.3 Framework Overview

Figure 7.2 shows the proposed framework. Starting from an ***①Input Sensor***, two main computation streams are constructed: ***②Tracking & Reconstruction***

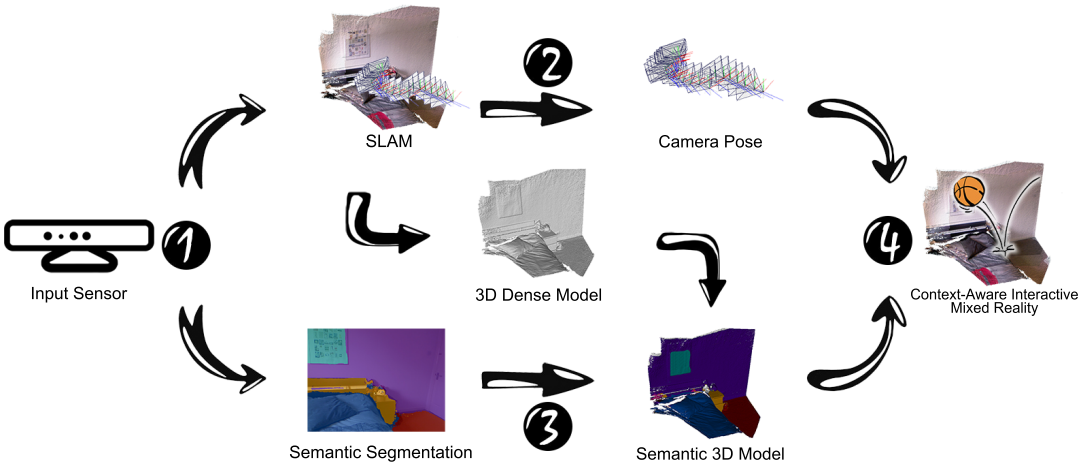


Figure 7.2: Flowchart demonstrates the whole framework.

Stream and ③***Context Detection & Fusion Stream***, which are finally merged and output to the ④***Interactive MR Interface*** for generating context-aware virtual-real interactions.

7.3.1 Input Sensor

An *input sensor*, a RGB-D camera such as Microsoft Kinect, ASUS Xtion series or built-in sensors on Microsoft HoloLens, is used to acquire the depth information directly for the 3D reconstruction of the environment. Monocular or stereo cameras would also work if combined with dense SLAM systems (Newcombe et al. 2011b), but the accuracy and real-time performance of Mono devices are not guaranteed.

7.3.2 Camera Tracking & Reconstruction Stream

The *tracking & reconstruction stream* shown in the upper path of Figure 7.2 processes the video captured by the input sensor. A SLAM system continuously estimates the camera pose while simultaneously reconstruct a 3D dense model. This is a typical method used in latest MR systems such as Microsoft HoloLens for implementing geometry-aware MR. A dense 3D model is served as the spatial collision mesh and the inverse of the camera pose extracted from SLAM guides the movement of the collision mesh to visually correspond to the real-world objects.

7.3.3 Context Detection & Fusion Stream

The lower path of Figure 7.2 shows the *Context Detection Stream*. 2D image sequences from the input sensor are context sources to be processed by semantic segmentation algorithms that can densely output the pixel-wise object attributes and properties of the scene. Based on the semantic segmentation information, the context information relevant for implementing context-aware experience is generated. Then the 2D semantic segmentation results are projected onto the scene and fused with the 3D dense model (from *tracking & reconstruction stream*) to generate a semantic 3D model based on the camera pose of the current frame.

7.3.4 Interactive MR Interface

The semantic 3D model are combined with the camera pose to provide a context-aware MR environment. High-level interactions can be designed based on the semantics. Furthermore, tools can be developed to facilitate design and automatic constructions of complex MR interactions in different applications.

The advantages of the proposed framework are:

1) **Accurate 3D Semantic Labeling**: The Context Detection & Fusion Stream can predict a pixel-wise segmentation of the current frame, which is further fused onto the 3D dense model. The semantic 3D model is generated with each voxel contains the knowledge of the context information of the environment. The voxel-based context-aware model delivers the semantic information through ray-cast queries about the object properties in order to generate different interactions. Object properties can be high-level descriptions, for example types of material and interaction attributes.

2) **Real-time Performance**: In deep learning based approaches the semantic segmentation is computationally expensive especially for processing frame by frame in real-time applications. We achieve the real-time performance by storing the semantic information into the 3D dense model after the initial segmentation process, so that the semantic segmentation is not processed at every frame, but at certain

frames.

3) **Automatic Interaction Design**: With the context information available, virtual and real object interactions can be designed and computed by feeding the object attributes of the real world to the target software module for processing e.g. a physics module or an agent AI module. For example, realistic physical interactions between virtual and real objects can be computed by feeding the material properties of the real world to physics simulation algorithm (such as our throwing plates game in the following section).

7.4 implementation

We present our novel MR framework in the context of object material-aware interactions as an implementation example to demonstrate the concept of context-aware MR. The material properties in the MR generates realistic physical interactions based on the objects' material property. This example implementation is also used for accuracy study and user experiment.

7.4.1 Camera Tracking and Model Reconstruction

The accurate camera tracking and dense 3D model reconstructions of the environment are achieved by adopting a dense SLAM system (Newcombe et al. 2011a), which estimates camera poses and reconstructs the 3D model in real-time. Depth images from a Kinect sensor are projected into the 3D model. The camera pose and a single global surface model can be simultaneously obtained through a coarse-to-fine iterative closest point (ICP) algorithm. The tracking and reconstruction processes consist four steps:

1) Each pixel acquired by the depth camera is firstly transformed into the 3D space by the camera's intrinsic parameters and the corresponding depth value is acquired by the camera;

2) An ICP alignment algorithm is performed to estimate the camera poses between the current frame and the global reconstructed model;

3) With the available camera poses, each consecutive depth frame can be fused incrementally into one single 3D reconstruction by a volumetric truncated signed distance function (TSDF);

4) Finally, a surface model is predicted via a ray-casting process.

A Microsoft Kinect V2 is used as the input sensor with OpenNI2 driver to capture RGB images and calibrated depth images at the resolution of 960x540 at 30 frames per second.

7.4.2 Deep Learning for Material Recognition

We trained a deep neural network for the 2D material recognition task. Our neural network is implemented in *caffe* (Jia et al. 2014) based on the CRF-RNN architecture (Zheng et al. 2015), which combines the FCN with Conditional Random Fields (CRF) based on the probabilistic graphical modelling for contextual boundary refinement. We use the Materials in Context Database (MINC) (Bell et al. 2015) as the training database that contains 3 million labelled point samples and 7061 labelled material segmentations in 23 different material categories.

The VGG-16 pre-trained model for ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) (Simonyan and Zisserman 2015) is used as the initial weights of our neural network. Based on the MINC dataset, we then fine-tune the network from 1000 different classes (ImageNet contains 1000 classes of labels) to 23 class labels as the output. VGG-16 is a CNN model specifically designed for classification tasks and only produces a single classification result for a single image. Therefore, we manually cast the CNN into a Fully Convolutional Network (FCN) for pixel-wise dense outputs (Shelhamer et al. 2017). By transforming the last three inner product layers into convolutional layers, the network can learn to make dense predictions efficiently at the pixel level for tasks like semantic segmentation. The fully-connected CRF model is then integrated into FCN to improve the semantic labelling results.

Fully-connected CRF encodes pixel labels as random variables form a Markov Random Field (MRF) (Kindermann and Snell 1980) conditioned on a global observation (the original image). By minimising the CRF energy function in the Gibbs

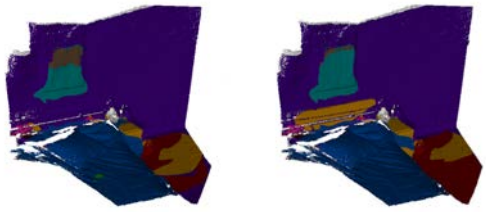
distribution (Ladick' et al. 2009), we obtain the most probable label assignment for each pixel in an image. With this process, the CRF refines the predicted label using the contextual information. It is also able to refine weak semantic label predictions to produce sharp boundaries and better segmentation results (see Figure 7.8 for the comparison of FCN and CRF-RNN). During the training process, CRF is implemented by multiple iterations, each takes parameters estimated from the previous iteration, which can be treated as a Recurrent Neural Network (RNN) structure (Zheng et al. 2015).

As the error of CRF-RNN can be passed through the whole network during a backward propagation, the FCN can generate better estimations for CRF-RNN optimization process during the forward propagation. Meanwhile, CRF parameters, such as weights of the label compatibility function and Gaussian kernels can be learned from the end-to-end training process.

We use 80% of the 7061 densely labelled material segmentations in the MINC dataset as the training dataset and the rest of 20% as testing sets. The training dataset is trained using a single Nvidia Titan X GPU for 50 epochs, after which there is no significant decrease of loss. For testing results, we obtain a mean accuracy of 78.3% for the trained neural network. The trained network runs at around 5 frames per second for the 2D dense semantic segmentation at the resolution of 480x270 on a Nvidia Titan X GPU. We input 1 frame into our neural network for every 12 frames according to our test to achieve a trade-off between the speed and accuracy.

7.4.3 Bayesian Fusion for 3D Semantic Label Fusion

The trained neural network for material recognition only infers object material properties in the 2D space. As the camera pose for each image frame is known, we can project the semantic labels onto the 3D model as textures. A direct mapping can cause information overlapping, since accumulated weak predictions and noises can lead a bad fusion result as shown in Figure 7.3 (a), where boundaries between different materials are blurred. We solve this issue by utilising the dense pixel-wise semantic probability distribution produced by the neural network over every class.



(a) (b)

Figure 7.3: (a) 3D semantic label fusion using direct mapping. (b) 3D Semantic Label Fusion with Bayesian fusion.

Therefore, we are able to improve the fusion accuracy by projecting the labels with a statistical approach using the Bayesian fusion (Armeni et al. 2016) (Hermans et al. 2014) (Zhao et al. 2017) (McCormac et al. 2017). Bayesian fusion enables us to update the label prediction l_i on 2D images I_k within the common coordinate frame of the 3D model.

$$P\left(x = l_i | I_{1,\ldots,k}\right) = \frac{1}{Z} P\left(x = l_i | I_{1,\ldots,k-1}\right) P\left(x = l_i | I_k\right) \tag{7.1}$$

where Z is a constant for the distribution normalization. The label of each voxel is updated with the corresponding maximum probability $p\left(x_{max} = l_i | I_{1,\ldots,k}\right)$. The Bayesian fusion guides the label fusion process and ensures an accurate mapping result over the time to overcome the accumulated errors to some extent. Figure 7.3 (a) shows without the Bayesian fusion, the label fusion results are less clear due to the overlapping of weak predictions. In contrast, 7.3 (b) with the Bayesian fusion, the fusion results are much cleaner.

After semantic information fusion into the 3D model, we can get a 3D semantic labelled model. Although the Bayesian fusion is used to guide the fusion process, due to accumulation of the 2D segmentation error and the tracking error, in some area, the semantic information still not perfectly matches the model structure (see Figure 7.4). Next we explain how to further improve the fusion accuracy by proposing a new CRF label refinement process on 3D structures.

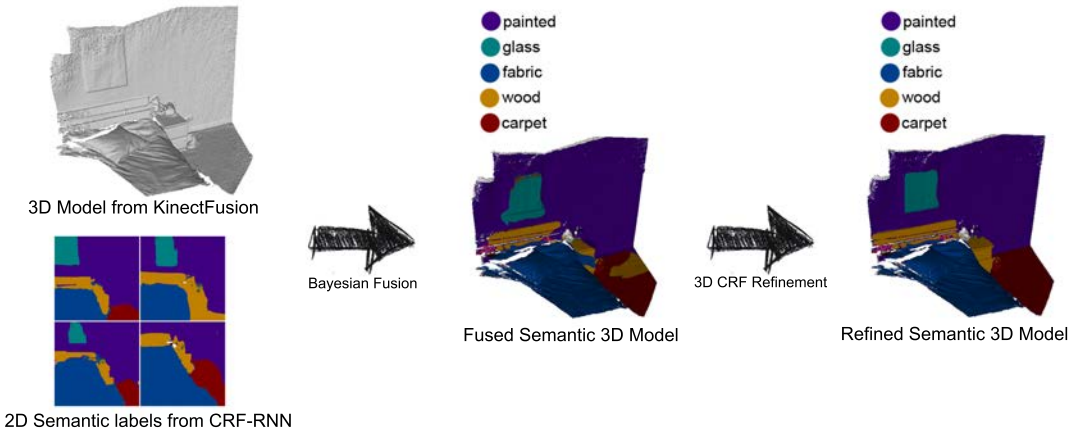


Figure 7.4: 3D semantic label fusion and refinement.

7.4.4 3D Structural CRF Label Refinement

We further improve the accuracy of the 3D labelling with a final refinement step on the semantic information using the structural and color information of the vertices of the 3D semantic model. From the fully connected CRF model, the energy of a label assignment x can be represented as the sum of unary potentials and pairwise potentials over all i pixels:

$$E(x) = \sum_i \psi_u(x_i) + \sum_i \sum_{j \in N_i} \psi_p(x_i, x_j) \tag{7.2}$$

where the unary potential $\psi_u(x_i)$ is the cost (inverse likelihood) of the i_{th} vertex assigning with the label x. In our model implementation, we use the final probability distribution from the previous Bayesian Fusion step as the unary potential for each label of every vertex. The pairwise potential is the energy term of assigning the label x to both i_{th} and j_{th} vertices. We follow the efficient pairwise edge potentials in (Krähenbühl and Koltun 2011) by defining the pairwise energy term as a linear combination of Gaussian kernels:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{M} w^{(m)} k_G^{(m)}(f_i, f_j) \tag{7.3}$$

where w^m are the weights for different linear combinations, k_m^G are m different Gaus-

sian kernels that f_i and f_j correspond to different feature vectors. Here, besides the commonly used feature space in (Krähenbühl and Koltun 2011) (Zheng et al. 2015) such as the color and the spatial location, the normal direction is also considered as a feature vector to take the full advantage of our 3D reconstruction step:

$$\begin{aligned} k_G\left(f_i, f_j\right) =& w^{(1)} exp\left(-\frac{\left|p_i - p_j\right|^2}{2\theta_p^2}\right) \\ &+ w^{(2)} exp\left(-\frac{\left|p_i - p_j\right|^2}{2\theta_{pI}^2} - \frac{\left|I_i - I_j\right|^2}{2\theta_I^2}\right) \\ &+ w^{(3)} exp\left(-\frac{\left|p_i - p_j\right|^2}{2\theta_{pn}^2} - \frac{\left|n_i - n_j\right|^2}{2\theta_n^2}\right) \end{aligned} \tag{7.4}$$

where p_i and p_j are pairwise position vectors; I_i and I_j are pairwise RGB color vectors; n_i and n_j are pairwise normal directional vectors. The first term is the smoothness kernel assuming that the nearby vertices are more likely to be in the same label, which can efficiently remove small isolated regions (Shotton et al. 2009)(Krähenbühl and Koltun 2011). The second term represents the appearance kernel that takes into account of the color consistency, since the adjacent vertices with similar color(s) are more likely to have the same label. The third term is the surface kernel which utilizes the 3D surface normal as a feature that vertices with similar normal directions are more likely to be the same label.

By minimizing Equation 7.2, semantic labels on our 3D model are further refined according to the color and the geometric information, which can efficiently eliminates the "label leaking" problem caused by the 2D semantic segmentation errors and the camera tracking errors (see Figure 7.4).

7.4.5 Interaction Interface

A user interface is developed with two layers. The top layer displays the current video stream from a RGB-D camera, whilst the semantic 3D model serves as a hidden physical interaction layer to provide an interaction interface. In the interactive MR application, a virtual camera is synchronized with predicted camera poses for

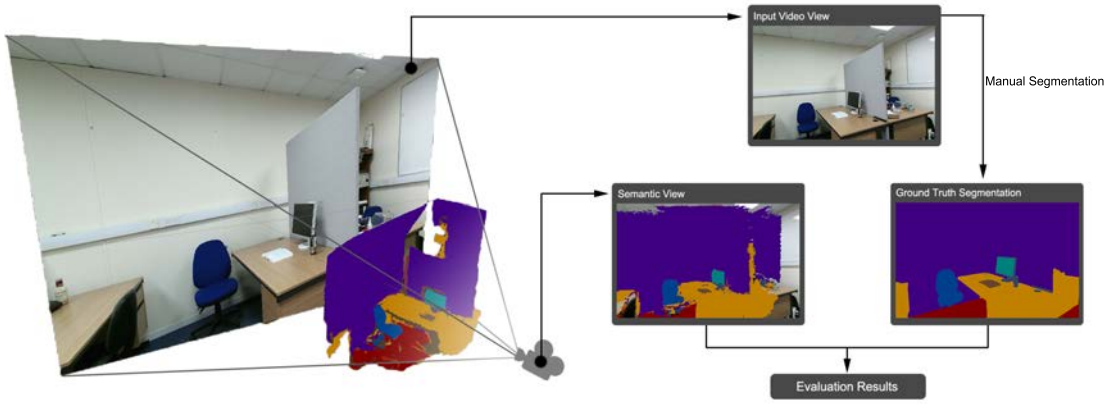


Figure 7.5: The evaluation framework

projecting the 3D semantic model onto the corresponding view of the video stream. Figure 7.5 shows that the back layer of the interface displays the video stream feed from the RGB-D camera; A semantic interaction 3D model is in front of the video layer for handling interactions of different materials (Green: glass, Purple: painted, Blue: fabric, Yellow: wood, Red: carpet). The virtual and real physical interactions are performed on the interaction model. The context-aware interaction model is invisible to allow users interact with the real-world objects to experience an immersive MR environment. The interaction layer also computes real-time shadows to make the MR experience even more realistic. An oct-tree data structure accelerates the ray-casting queries for the material properties to improve the real-time performance. Finally, corresponding physical interactions based on semantic information e.g. different materials are achieved through physics simulations.

7.5 Example Applications

Based on our implementation, two FPS games are developed to demonstrate the concept of the proposed material-aware interactive MR. Next, we describe the design of the interactions and evaluations.

Games are interaction demanding applications driven by computational performance and accurate interactions in virtual space. We designed two MR game that

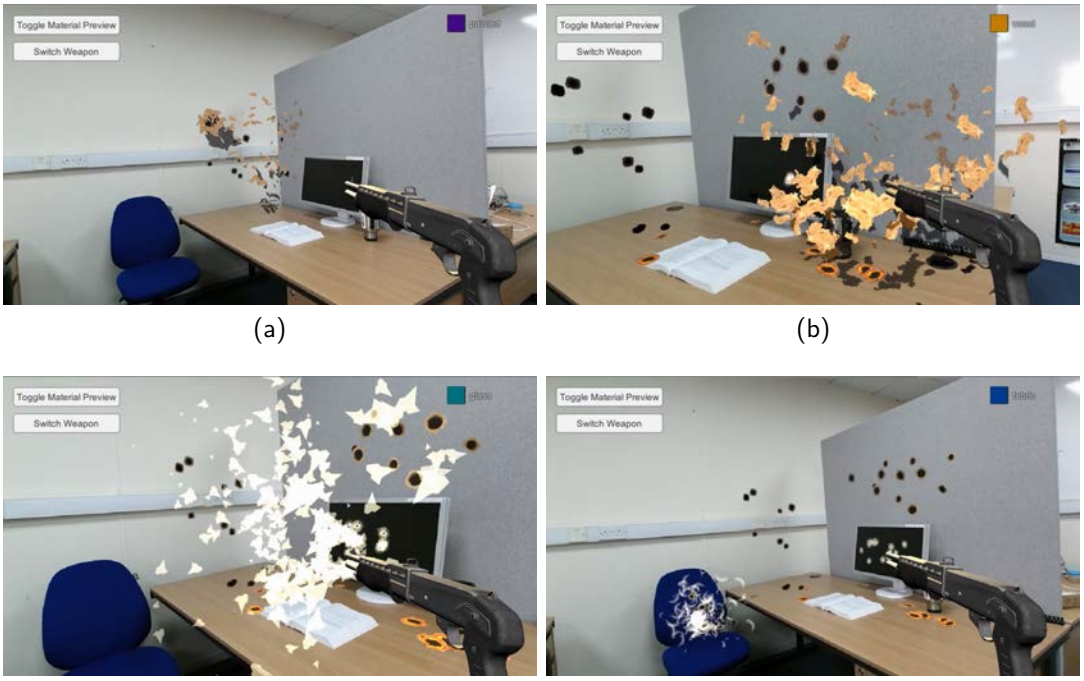


Figure 7.6: The screenshots of our MR shooting game. The interaction is different when shooting (a)walls, (b)desks, (c)computer screen and (d)chair.

can directly interact with the real-world objects. A shooting game is designed to show material-aware interactions between bullets and the real world objects. The shoot scenario is chosen, because we want to test the accuracy of the semantic 3D model using ray-cast queries. In this game, as shown in Figure 7.6, multiple interactions for different materials have been implemented including different bullet holes, flying chips and hitting sound when hitting different objects: (a)walls, (b)desks, (c)computer screen and (d)chair. The interaction for different material context is as real as possible.

Another way to show the capability of the context-aware framework is to match the interaction results to the user's anticipation of the interaction results using everyday scenarios that familiar to users, i.e. testing the immersive experience of the MR system from the user's perspective. The second example is designed to match the user expectations for material-specific physical interactions.

As shown in Figure 7.7, users throw virtual plates onto real world objects of the

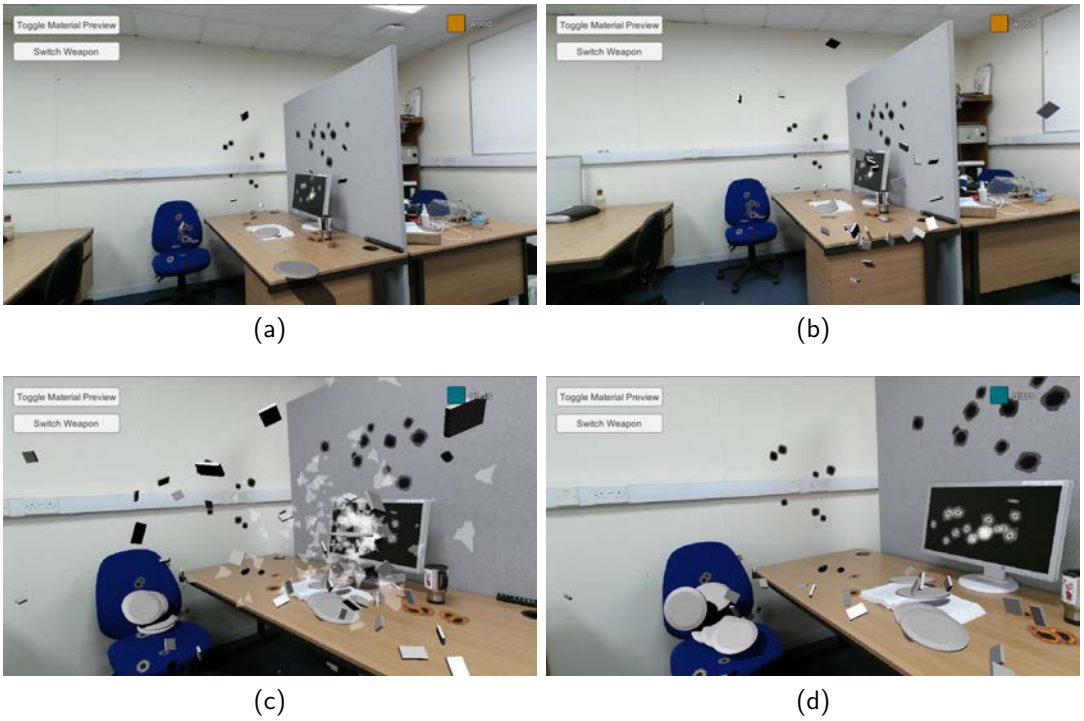


Figure 7.7: The screenshots of our MR throwing plates game. The interaction is different when throwing plates to (a)book, (b)desks, (c)computer screen and (d)chair.

MR environment, resulting material-aware physical interactions induced by various material properties of the real objects. In Figures 7.7 (a) and (b), virtual plates are broken when felling onto the desk, but bounced back when colliding with a book; in (c) when colliding with a computer screen, the plate is broken with the flying glass chips; in (d), the plate remain intact colliding with a soft chair.

7.6 Experimentation

7.6.1 Accuracy Study

Multiple factors affect the accuracy of the system: (1) the camera tracking, (2) the 3D model reconstruction, (3) the deep semantic material segmentation, (4) the 2D to 3D semantic model fusion and (5) the implementation of ray-casting. Therefore, it is extremely difficult to evaluate the accuracy of every single part of the system, separately. We conducted an end-to-end accuracy study to directly evaluate the

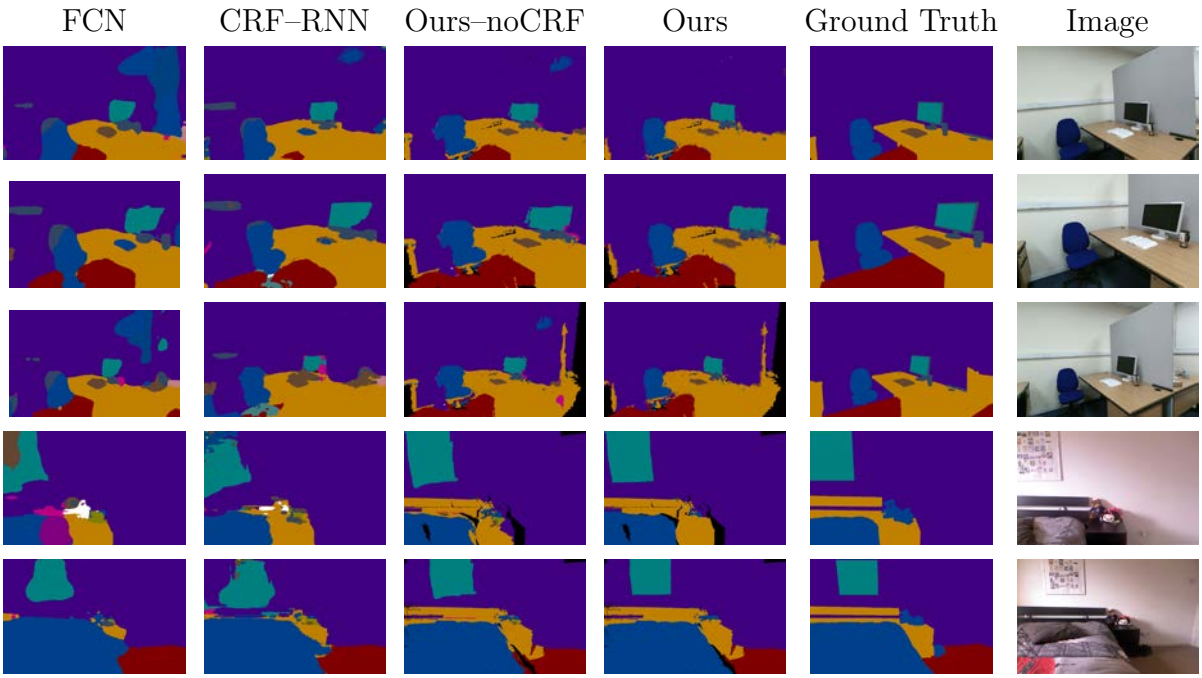


Figure 7.8: Semantic segmentation samples for each column from left to right: FCN, CRF-RNN, Ours without CRF refinement, Ours with CRF refinement, Ground Truth, Input image

accuracy of the dense ray-casting queries of the 3D semantic model, because it directly determines the accuracy of interactions. A total of 25 key-frames from two different scenes (office and bedroom) are selected, and at the same time, the 2D projections of the 3D semantic models are captured as the dense ray-casting query results at the corresponding key-frames (see Figure 7.9). Ground truth for the accuracy evaluation is obtained by manually labelling 25 RGB images with the same material labels. The four common evaluation criteria (Shelhamer et al. 2017) (Zheng et al. 2015) for semantic segmentation and scene parsing evaluations are used to evaluate the variations of pixel accuracy and region intersection over union (IU).

1. pixel accuracy $\frac{\sum_i n_{ii}}{\sum_i t_i}$
2. mean accuracy $\frac{1}{n_{c1}} \sum_i \frac{n_{ii}}{t_i}$
3. mean IU $\frac{1}{n_{c1}} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$

4. frequency weighted IU $\frac{1}{\sum_k t_k} \sum_i \frac{t_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$

where n_{ij} represents the the number of pixels of class i predicted to be class j; n_{c1} is the total number of classes; $t_i = \sum_j n_{ij}$ is the total number of pixels of class i.

As can be seen from Table 7.1, after 2D-3D fusion, 3D refinement and finally 3D-2D projections, our framework can provide more accurate semantic segmentation results compared with the 2D methods such as FCN and CRF-RNN from the metrics of pixel accuracy and mean accuracy. Our method also produce less error labels compared with other, which can be revealed from mean IU and frequency weight IU. Figure 7.8 shows some semantic segmentation samples. Taking the advantages of the 3D constraints and refinement in our proposed framework, our semantic segmentation results are more uniform, sharp and accurate.

Table 7.1: Accuracy study results compared with other 2D semantic segmentation algorithms

	pixel acc.	mean acc.	mean IU	f.w. IU
FCN(Shelhamer et al. 2017)	81.61	63.69	49.54	76.16
CRFRNN(Zheng et al. 2015)	85.68	51.73	41.32	79.76
Ours_noCRF	**87.86**	**70.69**	**54.81**	**81.86**
Ours	**89.42**	**72.06**	**56.32**	**84.30**

7.6.2 User Experience Evaluation

We conduct a user study to evaluate the effectiveness of the semantic-based MR system. Using the throwing plate game, three test conditions are designed by setting different collision responses:

1) *No Collision Mesh*: Virtual plates were thrown into the real world without any collision being detected.

2) *Uniform Collision Mesh*: Virtual plates interact with the real world with the uniform collision mesh being activated, but no object-specific interaction is generated.

3) *Semantic Collision Mesh*: Physics responses of the virtual plates with the real-world objects are dependent on the material properties of the objects in the real world.

The objective of the user study is to assess the realism of the MR environment by measuring how much the semantic-based interactions matches the user anticipations. We investigate whether or not the semantic-based interactions can significantly improve the realism of MR systems and delivers immersive user experience.

Firstly, we evaluate the realism of the physical interactions such as collision responses in MR systems. We test to see if users are able to detect differences in these three interaction conditions between virtual and real objects in short video clips, and whether or not the realism in MR can be improved via context-aware physical responses. Secondly, to ensure the quality of qualitative study, we test if there is any risk that the user experience of the proposed MR system could be affected by his or her previous engagement with computer games and MR or VR technologies.

7.6.2.1 Participants

A questionnaire was designed and an online servery platform was used to conduct the user study. Anonymous participants were recruited without restrictions on age and gender. Each participant was firstly asked whether he/she had any previous experience with FPS games and VR/AR games, and then asked to watch the three video clips carefully. Each video clip can be viewed repeatedly, so that the participant can take time to digest and answer the questions. Each of the video clip was rated by participants on the scale of 1 (very bad) to 10 (very good) based on the quality of the MR interactions and realism.

The questionnaires are shared among students' chatting group. A total of 68 questionnaires were received, in which 6 responses were removed for reasons either participants did not confirm that they have watched the videos carefully or their viewing time was too short (less than 10 seconds) indicating little interest from the participants. Among the 62 valid questionnaires, 69.57% had experience with FPS

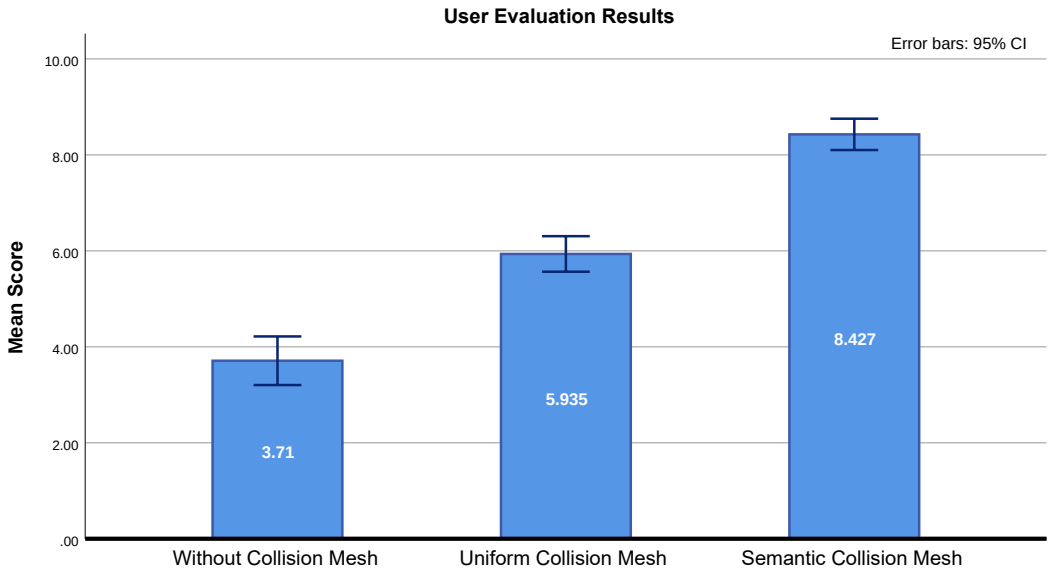


Figure 7.9: The user experience evaluation results.

games, 65.22% had experience with VR/AR games.

7.6.2.2 Results

We used the score from 1 to 10 as the interval data so that we can use parametric ANOVA to analysis the data. We have performed a repeated measure ANOVA test to analyze scores obtained for the three conditions. Mauchly's test indicated that the assumption of sphericity had been violated ($X^2 = 42.029, p < 0.001, \varepsilon = 0.665$). Therefore, we used Greenhouse-Geisser to correct the degrees of freedom. Main effects were found within the three conditions ($F_{1.330,81.135} = 212.293, p < 0.001$).

The following post hoc Bonferroni pairwise comparisons show that the *Semantic Collision Mesh* ($M = 8.427, SD = 1.995$) is significantly better than the other two MR conditions ($p < 0.001$), indicating that the proposed semantic interactions through the inference of material properties can greatly improve the realism of the MR system. We also found that the *Uniform Collision Mesh* ($M = 5.935, SD = 1.458$) offers much better MR experience ($p < 0.001$) than the *No Collision Mesh* ($M = 3.71, SD = 1.99$) but less realistic compared with the semantic context-aware MR. The mean scores of the three system conditions are shown in Figure 7.9.

Furthermore, as a large number of our participants have either experienced FPS games (69.57%) or VR/AR games (65.22%), we also conducted a between-subjects repeated measure ANOVA test to reveal whether this experience has an influence on the user when assessing the results due to prior exposure to VR, MR, and games technologies. It has been shown that the final test results are not affected by either the experience of FPS games ($p = 0.793$) or VR/AR games ($p = 0.766$).

7.6.3 User Feedback

Many participants were interested in the MR system and gave very positive comments and feedback about their MR experience that the system provides. Comments are such as *"This game (throwing plates) is amazing! I never experienced such MR experience before."*; some people commented on the importance of material-specific interactions even the low quality models, textures and animation being used in the current prototype, *"Although the interaction sometimes is not very obvious, it really makes a lot of difference."*; some people criticized the MR system without semantic interaction: *"The next second when the plates break when hitting a soft chair (indicating the MR system with Uniform Collision Mesh), I won't play it again, as it violates the basic physical law."*, while other people cannot wait to play our semantic interaction MR game *"Your game creates a realistic interactive experience, nice work! When will you release your game?"*.

7.7 Conclusion and Discussion

We show how deep semantic scene understanding methodology combined with dense 3D scene reconstruction can build high-level context-aware highly interactive MR environment. Recognizing this, we implement a material-aware physical interactive MR environment to effectively demonstrate natural and realistic interactions between the real and the virtual objects. Our work is the first step towards the high-level interaction design in MR. This approach can lead to better system design and evaluation methodologies in this increasingly important technology field.

There are some immediate directions for future research and we mention two such directions now. Although in this chapter we focus our discussions on material understanding and its semantic fusion with virtual scene in MR environment, the concept and the framework presented here are applicable to address many other context-aware interactions in MR, AR and even VR. The framework can be extended by replacing the training dataset with more general object detection databases for constructing different interaction mechanisms and context. Realistic physically-based sound propagation and physicall-based rendering using the proposed context-aware framework for MR are promising directions to pursue. Integrated with multi-modal interactions, immersive experience can be achieved. Our results have hinted that the study of semantic constructions in MR as a high-level interaction design tool is worth pursuing, as more comprehensive methodologies emerging, complex rich MR applications will be developed in the near future.

We believe that AI is not only for autonomous tasks of machines and robots, but also is for the improvement of human decision making when interacting with the real world through virtual interactions.

Chapter 8

Increase Tracking and Reconstruction Robustness – Learning-based Image Smoke Removal

8.1 Introduction

Surgical smoke is the by-product of using energy-generating devices that raise intracellular temperatures during surgery. Surgical smoke in intraoperative imaging and image-guided surgery (Tsui et al. 2013) can severely deteriorate the image quality (Weld et al. 2007) and pose hazards to surgeons (Planteféve et al. 2016). Therefore, improving the quality of intraoperative images is highly desirable for many clinical applications. Surgical smoke also poses significant issues (Ulmer 2008) in future advanced medical imaging tasks such as robotic surgery, real-time surgical reconstruction and Augmented Reality, where the effectiveness of computer vision is pertinent.

Physical smoke evacuation devices are available for removing surgical smoke, but such devices hardly usable for image-guided surgery. Recently research approaches

based on more conventional image processing methods were proposed for filtering out the smoke, then trying to recover images as sharply and clearly as possible (Kotwal et al. 2016) (Baid et al. 2017) (Luo et al. 2017) (Tchaka et al. 2017) (Wang et al. 2018).

But these methods will always over-enhance the image and suffer from fidelity loss problem. More recently, end-to-end deep learning (DL) approaches (Cai et al. 2016) have been introduced to solve the de-hazing and de-smoking problems, which achieve promising results. However, there are still several challenges need further before it can be introduced to medical practice:

- It is extremely difficult to collect large amounts of data for the effective learning of the implicit de-smoking function, especially for surgical scenes.
- Learning-based methods can be over-fitted to the limited amount and variation of training data, leading to unstable performance when testing on real-world data.
- Smoke can be also regarded as an important signal of the ablation process. Removing the smoke sometimes can have the reverse effect if it is not quantifiable and controllable.

In this chapter, we formulate the smoke detection and removal as jointly learning processes and propose a novel computational framework for unsupervised collaborative learning of smoke detection and removal from rendering smoke on laparoscopic videos. The contribution of this work include:

- We innovatively integrate a render engine into our learning framework for continuously outpouring unlimited training data without any manual labeling needed.
- We decompose the smoke removal task into two loosely-coupled sub-module tasks: pixel-level smoke detection and smoke removal based on the detection results, which not only can prevent over-fitting to synthetic data but can also make the surgeon aware of how much smoke is removed.

- We propose a novel Generative Collaborative Networks (GCN) training framework that maximally exploits the potential of our smoke detection and removal networks.

- The quantitative and qualitative evaluation results prove that our proposed method outperforms the GAN framework and the state-of-the-art smoke removal approaches.

Compared to conventional image processing approaches, our proposed framework is able to remove smoke with a global contextual understanding and recover more realistic tissue colour. Although trained on synthetic images, the experimental results show that our network is able to effectively remove smoke on laparoscopic images with real surgical smoke.

8.2 Related Work

Image de-hazing and de-smoking are the tasks that have been researched for decades in the image processing and computer vision communities for recovering clear outdoor scenes affected by bad weather. Typical methods for smoke removal are either based on image processing or machine learning.

8.2.1 Atmospheric Scattering Model

One of the most classic model to describe the hazy or smoky image is the atmospheric scattering model(McCartney and Hall 1977) (Narasimhan and Nayar 2003) (Nayar and Narasimhan 1999)

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{8.1}$$

where $I(x)$ is the observed hazy image, $J(x)$ is the clear scene to be recovered, A is the global atmospheric light, $t(x)$ is the medium transmission, which can be described by $t(x) = e^{-\beta d(x)}$, where β is the atmosphere scattering coefficient and

$d(x)$ is the distance. However, the atmospheric scattering model is based on the strong assumption that the haze is homogeneous, and the light source is parallel and even beams (such as sunlight). Unfortunately, for the surgical smoke in MIS, smoke concentration can vary greatly, so $t(x)$ is very hard to predict. The light source is usually uneven due to the very close distance between light and tissues.

8.2.2 Dark Channel Prior based de-smoking

The dark-channel prior (DCP) proposed by He et al (He et al. 2011) is a simple but effective approach for predicting transmission map based on the observation of the natural property of haze-free images – pixels should have at least one color channel with very low-intensity values. However, for real-world images, it will cause chromatic change and fidelity loss. Also, for MIS scenes, the close-distance direct light source will produce over-illuminated pixels such as reflection of tissues, light color fat tissues, which will violate the DCP assumption and be falsely detected as hazy.

Tchaka (Tchaka et al. 2017) adopted adaptive DCP with histogram equalization to remove smoke for endoscopic images and improve the recovered image quality. However, several parameters are chosen empirically, and due to the limitation of DCP, although histogram equalization can enhance the color and contrast, the original and real color will not be preserved.

8.2.3 Optimization-based De-smoking

Fattal et al (Fattal 2008) further refined the model in Equation 8.1, and take the surface shading into account in addition to the scene transmission and use Gaussian Markov Random Field (MRF) model to recover the haze-free image. Similarly, Nishino (Nishino et al. 2011) modeled chromaticity and depth using factorial MRF for more accurate scene radiance estimation. Tan et al (Tan 2008) proposed a local contrast maximizing method based on the observation that hazy-free images tend to have much higher contrast, which also relies on optimizing MRF models.

Meng et al (Meng et al. 2013a) introduced an inherent boundary constraint on the transmission function that can recover more image details and structures. Baid et al (Baid et al. 2017) presented a unified Bayesian formulation for joint de-smoking, specularity removal, and de-noising in laparoscopy images. They propose several priors by probabilistic graphical models and sparse dictionaries to model the image color and texture of the un-corrupted image. The variational Bayes Expectation Maximization(VBEM) optimization is used to minimize the overall energy function and infer the un-corrupted images from corrupted images.

However, although the above MRFs priors are well-designed, these hand-crafted prior models have limited expressive power, and lack a global contextual understanding for an ill-posed problem like de-smoking. Another common weakness of these methods is that they were all trying to minimize the prior features that tend to be hazy, which usually lead to over-enhanced color and contrast and also suffer from fidelity loss.

8.2.4 Learning based De-smoking

With the recent success of deep learning, many deep learning frameworks are introduced to solve the de-hazing and de-smoking problems. DehazeNet (Cai et al. 2016) is an end-to-end learning system for single image haze removal by learning the medium transmission map, which is subsequently used to recover a haze-free image by the atmospheric scattering model. Similarly, AOD-Net (Li et al. 2017a) integrate the numerical computation of atmospheric scattering model into the network structure and achieve all-in-one end-to-end training. However, these networks are still based on the Atmospheric Scattering Model that is not suitable for MIS scenes, also the network structures are very shallow and out-dated for learning and recovering fine details.

8.2.5 Novelity to previous work

Most of the above work rely on the Equation 8.1 to solve the de-hazing problem. However, in MIS scenes, most of the smoke is non-uniform, light beam is unparallel and uneven, making it an ill-posed problem. In this chapter, we reformulate the Equation 8.1 as fully end-to-end learning processes, by estimating the smoke mask first and use it as prior knowledge for another neural network to learn the ill-posed smoke removal function. We not only achieve better smoke removal results, but also reduce over-fitting and make the network more robust to real-world images. In addition, the pixel-level smoke detection result can lead to many useful applications such as smoke volume estimation and increase the awareness of surgical smoke for surgeons.

8.3 Methods

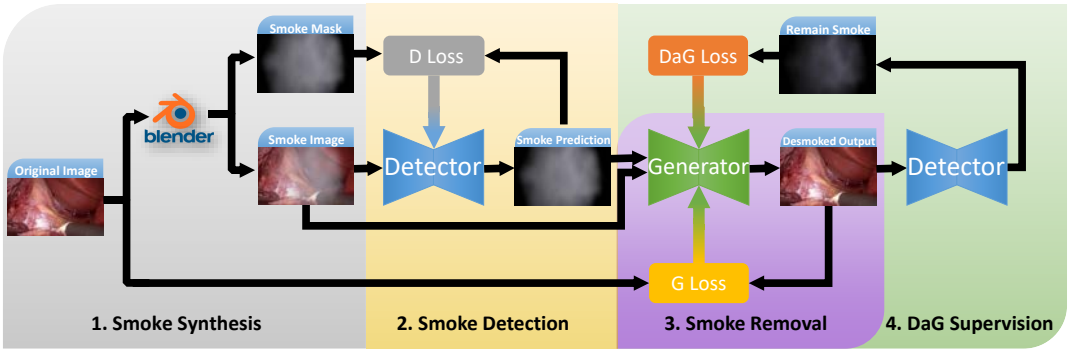


Figure 8.1: Overview of our framework for unsupervised learning of smoke removal

Our aim for removing smoke is very straightforward: We want to remove the smoke while maximally keep the non-smoke area as original color. Inspired by Equation 8.1, we decompose the smoke removal task into two sub-tasks: smoke detection and removal. Two fully convolutional networks are used to learn the smoke detection and removal separately but also cooperatively:

- Smoke detection network focus on detecting smoke and providing a pixel-level smoke mask

- Smoke removal network focus on remove smoke based on the smoke mask and smoke image
- Smoke detection network serves as a supervision to exam the smoke removal result and provides gradients for optimizing smoke removal network.

As can be seen from Figure 8.1, our training pipeline consists 4 main parts – 1. Smoke Synthesis, 2. Smoke Detection, 3. Smoke Removal, and 4. Detection-after-generation (DaG) supervision. Each of the 4 components will be introduced in details in the following subsections.

8.3.1 Smoke Synthesis

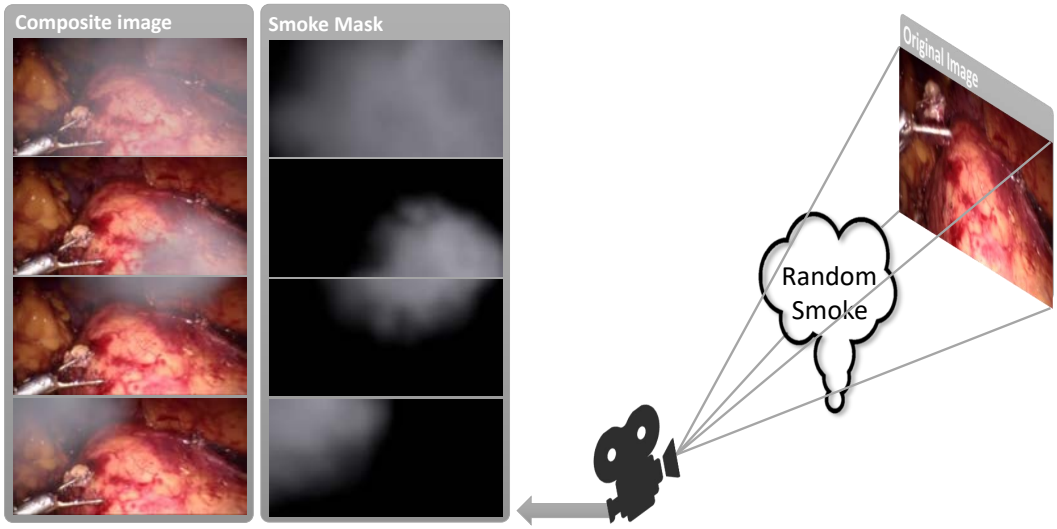


Figure 8.2: Left: the rendered images and smoke masks. Right: The 3D illustration of the rendering process.

Collecting large datasets for training neural networks is extremely costly and time-consuming, especially for medical datasets, which not only take up lots of valuable medical resources, but also require critical accuracy and great quantity to satisfy the standard to be used in medical practice. For the smoke detection and removal tasks, it is even more difficult as we require the pairs of images with/without the presence of smoke and also the smoke density mask, which are nearly impossible to acquire by human's manually labeling.

To tackle this problem, we employ a modern render engine for continuously render smoke on laparoscopic images to generate smoke images and smokes mask for the training of pixel level smoke detection and removal tasks. We use Blender [1] – an open source 3D creation software for the synthesis of smoke images for training.

The advantages of using a render engine instead of using physical haze formation model (Cai et al. 2016) (Tang et al. 2014) or Perlin noise (Bolkar et al. 2018) are: 1) in laparoscopic scenes, the surgical smoke is generated locally and independent of depth, so there is no meaning to use traditional haze model for rendering surgical smoke; 2) render engine can produce more realistic smoke shape and density variation based on the built-in realistic smoke rendering model.

The real laparoscopic images from Hamlyn Centre Laparoscopic / Endoscopic Video Datasets [2] (Ye et al. 2017) and Cholec80 Dataset [3] (Twinanda et al. 2017) are used as background images. The variance of the Laplacian (Bansal et al. 2016) are first used for screening images, and a second round manually inspection ensures the images are without the presence of surgical smoke to prevent confusing our networks. Totally 23,005 images are sampled from 91 videos as the smoke-free source images.

Smoke is rendered on each of the images with random intensity I_{rand}, density D_{rand} and position P_{rand} on the background image $I_{smoke-free}$ to simulate the image with surgical smoke I_{smoke}.

$$Blender(I_{smoke-free}, I_{rand}, D_{rand}, P_{rand}) \begin{cases} I_{mask} \\ I_{smoke} \end{cases} \tag{8.2}$$

The variation of rendered smoke ensures that our network will not over-fitting to certain smoke intensity, density and location. With the help of powerful render engine, we are able to synthesize unlimited amount of realistic images with the presence of simulated surgical smoke for training our network.

[1] https://www.blender.org/
[2] http://hamlyn.doc.ic.ac.uk/vision/
[3] http://camma.u-strasbg.fr/datasets

8.3.2 Smoke Detection

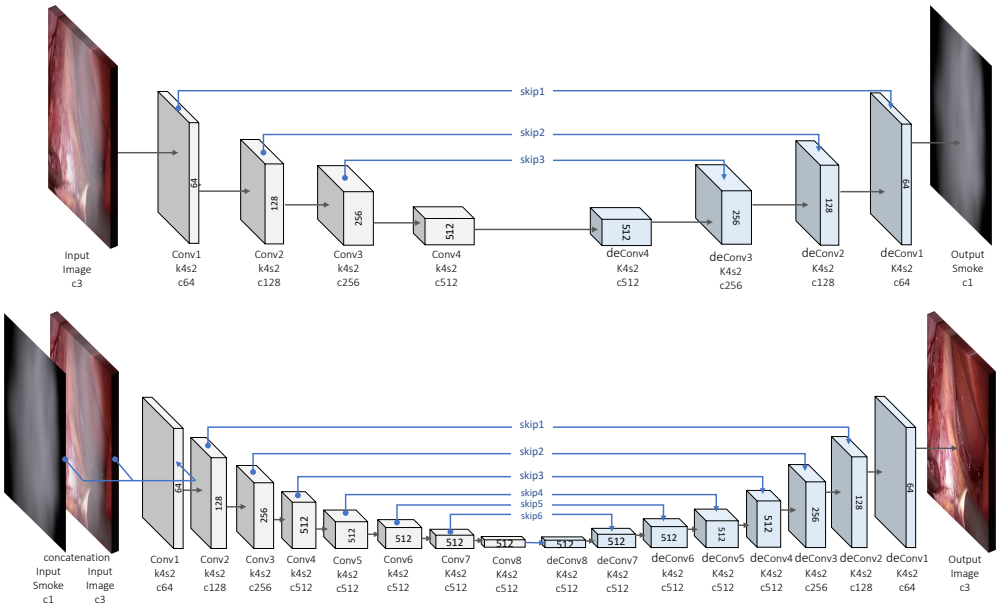


Figure 8.3: Network structures of smoke detection network (top) and smoke removal network (bottom)

We proposed to firstly use a smoke detection network to generate pixel-wise smoke density. The benefits of smoke detection are three-fold:

- Smoke detection provides a pixel-level smoke density for surgeon's awareness of the amount and position of the presence of surgical smoke.
- The detected smoke serves as a prior information inputted into the following smoke removal network.
- Smoke detection network is further used as a supervision to help optimize the smoke removal network. (see Section 8.3.4)

We employ a U-Net (Ronneberger et al. 2015) based fully convolutional encoder-decoder network structure with parameters θ_d for pixel level smoke detection:

$$D(I_{smoke}) \underset{\theta_d}{\rightarrow} I_{mask} \tag{8.3}$$

As shown in Fig. 8.3, our smoke detection network consists of 4 convolutional layers as an encoder to efficiently abstract the input image into a high-dimensional feature tensor with $1/2^4$ original size and 512 channels. For decoder, 4 symmetrical de-convolutional layers are used to recover the feature tensor into a full original size smoke mask. Each layer is with kernel size 4 and stride size 2, followed by leaky ReLU layers and batch normalization. Skip layers are connected with the corresponding layer pairs from encoder and decoder for preserving high-level information to ensure the high quality per-pixel smoke detection after up-sampling.

We intend to use a shallow network with fewer layers for the reason of:

- Smoke detection is a simple task compare to smoke removal, so a shallow network will be enough
- A shallow network will have fewer weights to prevent over-fitting to specific smoke patterns
- A shallow network will accelerate training and inferring speed.

The loss function for our smoke detection network is:

$$\begin{aligned}\mathcal{L}_D^{total} = \sum_{x,y}(\alpha_d \underbrace{\left|\widehat{I}_{mask}(x,y) - I_{mask}(x,y)\right|}_{L1\ loss} \\ + \underbrace{\beta_d \left|\widehat{I}_{mask}(x+1,y) - \widehat{I}_{mask}(x,y)\right|}_{x\ smooth\ term} \\ + \underbrace{\beta_d \left|\widehat{I}_{mask}(x,y+1) - \widehat{I}_{mask}(x,y)\right|}_{y\ smooth\ term})\end{aligned} \tag{8.4}$$

where $\widehat{I}_{mask}(x,y)$ and $I(x,y)_{mask}$ are the estimated smoke mask and ground truth smoke mask. We use a combination of an L1 loss and two smoothness terms for the total network loss. We take the L1 norms of the predict smoke masks' gradients along x and y directions as smoothness terms. Based on the fact that smoke tend to be smooth, applying penalty on the smoke masks' discontinuities can ensure accurate, smooth and realistic smoke mask prediction.

8.3.3 Smoke Removal

The smoke mask I_{mask} estimated by our smoke detection network is further used as an prior knowledge for learning smoke removal. As can be seen from the second network in Figure 8.3, the smoke mask I_{mask} and smoke image I_{smoke} are concatenated into a 4 channel layer before inputted into our smoke removal network G with parameters θ_g.

$$G(I_{mask} \oplus I_{smoke}) \underset{\theta_g}{\rightarrow} I_{smoke-free} \tag{8.5}$$

An encoder-decoder network similar to our smoke detection network is used for generating smoke-free images. We used a deeper network with 8 convolutional layers for the encoder to compress the input image into a 512 channel feature tensor and 8 de-convolutional layers to recover it into full-size smoke-free mask. To prevent the loss of details, following the U-Net structure (Ronneberger et al. 2015), skip connections are implemented for directly transferring high-level information to the bottom of the network. We used doubled number of layers for learning smoke removal as smoke removal is an ill-posed problem that require the contextual understanding of the image to recover the correct color of the regions covered by smoke.

The first part of the loss function of our smoke removal network is the L1 loss between the estimated smoke-free image and the original smoke-free image without simulated smoke:

$$\mathcal{L}_G^{L1} = \sum_{x,y} \left| \widehat{I}_{smoke-free}(x,y) - I_{smoke-free}(x,y) \right| \tag{8.6}$$

8.3.4 Detection after Generator (DaG) Supervision

To fully take advantage of our smoke detection network, we propose to use the smoke detection network as a second supervision to further guide the smoke removal process. The estimated smoke-free image $\widehat{I}_{smoke-free}$ is inputted into our smoke detection network after generated from our smoke removal network:

$$D(\widehat{I}_{smoke-free}) \underset{\theta_d}{\rightarrow} 0 \tag{8.7}$$

We want to make sure the smoke removal network G works well and there is no smoke left after, so our goal is to minimize the output of our detected smoke to provide gradients for our smoke removal network G. Therefore, the second part of the loss function is the L1 norm of the predicted smoke mask based on the predicted smoke-free image, which can also be expressed as the L1 norm of the detector after generator:

$$\begin{aligned} \mathcal{L}_G^{DaG} &= \sum_{x,y} \left| D(\widehat{I}_{smoke-free}(x,y)) \right| \\ &= \sum_{x,y} |D(G(I_{smoke}(x,y)))| \end{aligned} \tag{8.8}$$

Therefore, the total loss of our smoke removal network is:

$$\mathcal{L}_G^{total} = \alpha_g \mathcal{L}_G^{L1} + \beta_g \mathcal{L}_G^{DaG} \tag{8.9}$$

where α_g and β_g are the weights for $L1$ loss and DaG loss.

8.4 Experiments

In this section, we present our experiment setups and evaluation results of our proposed smoke detection and removal networks. We provide quantitative and qualitative comparisons of our results with 11 prior approaches.

8.4.1 Implementation details

Our networks are implemented in Tensorflow and trained on a workstation with a NVIDIA Titan X GPU (12G Graphic Memory).

For training, we apply the gradient descent steps of D and G separately to avoid interference between each other. Similar to the GAN training strategy, we firstly train D for one step and then train G for one step. When training G, we freeze the network parameters of D. Adam solver is used for training with hyperparameters learning rate 0.0002, and momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$, batch size

of 16. We empirically set the weights $\alpha_d = \beta_d = 1$, $\alpha_g = 1$, $\beta_g = 100$ based on several tests. In our implementation, drop-out is used in the 5th layer for our smoke detection network and 9th layer for our smoke removal network with the change of 50% to prevent overfitting.

The input images are re-sized to 256x256 pixel for efficient training and testing. For the total 23,005 image pairs of smoke-free images and rendered smoke images, 21000 images are randomly picked for training, and the remaining 2005 images are used for testing.

The training time is around 14 hours. When in testing mode, our networks can estimate smoke masks and smoke-free images at 45 frames per second.

8.4.2 Comparison Methods

For quantitative evaluation, we report a series of evaluation criteria in terms of the difference between the pair of smoke-free image and de-smoked results, including the Mean Squared Error (**MSE**), the Peak Signal-to-Noise Ratio (**PSNR** in dB) and the Structural Similarity Index (**SSIM**). The lower **MSE** and higher **PSNR** and **SSIM** mean that the estimated smoke-free images are similar to the real smoke-free images, which indicates a better de-smoking ability.

We compared our method with conventional de-smoking and de-haze methods including Dark Channel Prior (**DCP**) (He et al. 2011), Boundary Constraint and Contextual Regularization (**BCCR**) (Meng et al. 2013a), Fusion-based Variational Image Dehazing (**FVID**) (Galdran et al. 2016), Automatic Recovery of the Atmospheric Light (**ATM**) (Sulami et al. 2014), Color Attenuation Prior (**CAP**) (Zhu et al. 2015), DEnsity of Fog Assessment based DEfogger (**DEFADE**) (Choi et al. 2015), Enhanced Variational Image Dehazing (**EVID**) (Galdran et al. 2015), Non-local Image Dehazing (**NLD**) (Berman et al. 2017), Graphical Models and Bayesian Inference (**GMBI**) (Baid et al. 2017), and deep learning based methods including the All-in-One Dehazing Network (**AOD-NET**) (Li et al. 2017a), Image-to-Image Translation with Conditional Adversarial Networks (**PIX2PIX**) (Isola et al. 2017). All of the source codes are collected from authors' and third-parties' implemen-

tations with the default parameters in their papers. It is worth noticing that for DL-based methods (Li et al. 2017a) (Isola et al. 2017), we trained the networks with the same datasets and number of epochs with our networks for a fair comparison study.

8.4.3 Evaluation on Testing Dataset

We evaluated our trained model and compared with the previous approaches using our the testing dataset containing 2005 images. As can be seen from the box plots in Fig. 8.4 and Table 8.4.3, our method outperforms all of the previous de-hazing and de-smoking method in terms of MSE, PSNR and SSIM, with very small standard deviations, indicating the robustness of our system. We also report the averaged computational time for all of the compared methods in the last row. We can find that the DL-based methods take significantly less time to estimate smoke-free images compared to conventional methods. As our framework is a series-connection of two networks when testing, the computational time is doubled compared to the single network, but still can be running at 45 frames per seconds (1.5X real-time).

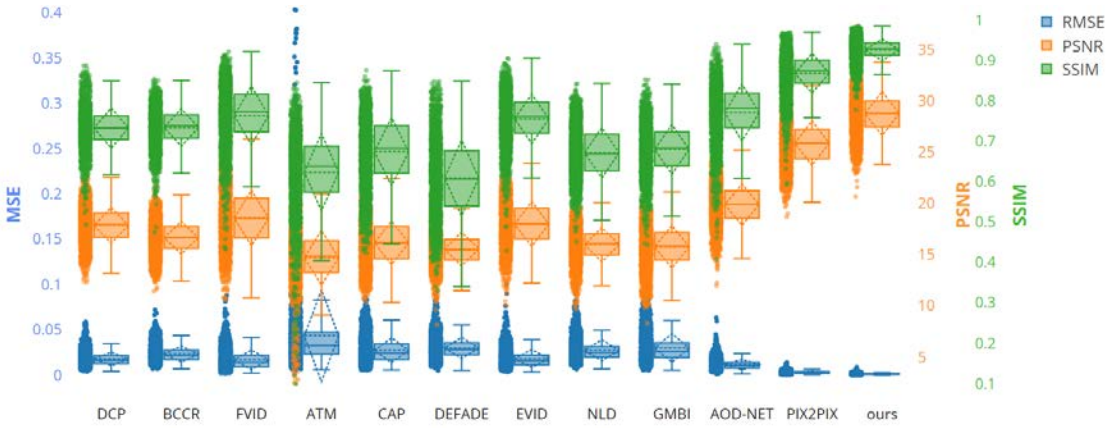


Figure 8.4: Box plots of the 3 metrics MSE, PSNR and SSIM for our result and 11 previous approaches.

As can be seen from Fig. 8.5, we display 6 sets of the smoke-free images $I_{smoke-free}$, smoke masks I_{mask}, rendered smoke images I_{smoke} (the only input to all methods), de-smoking results of 11 previous methods and the output of our

Table 8.1: QUANTITATIVE RESULTS

Method	DL?	Platform	Lower is better	Higher is better	Higher is better	Time/frame
			MSE	PSNR	SSIM	
DCP	No	Matlab	0.0178 ± 0.0075	17.8531 ± 1.7416	0.7328 ± 0.0498	3.6125
BCCR	No	Matlab	0.0231 ± 0.0088	16.6477 ± 1.5776	0.7344 ± 0.0465	0.2745
FVID	No	C/Matlab	0.0168 ± 0.0109	18.5761 ± 2.7109	0.7639 ± 0.0724	5.3597
ATM	No	Matlab	0.0433 ± 0.0506	14.7037 ± 2.6501	0.6236 ± 0.0958	21.5084
CAP	No	Matlab	0.0273 ± 0.0136	16.1510 ± 2.1195	0.6744 ± 0.0843	0.1175
DEFADE	No	Matlab	0.0299 ± 0.0116	15.5467 ± 1.6394	0.6092 ± 0.0936	2.1233
EVID	No	C/Matlab	0.0181 ± 0.0099	17.9780 ± 2.1788	0.7557 ± 0.0609	5.8055
NLD	No	Matlab	0.0273 ± 0.0117	15.9656 ± 1.6565	0.6718 ± 0.0624	5.0158
GMBI	No	Matlab	0.0298 ± 0.0157	15.7614 ± 2.0796	0.6802 ± 0.0615	2.2103
AOD-NET	Yes	Caffe	0.0115 ± 0.0057	19.8744 ± 1.9964	0.7711 ± 0.0704	0.0173
PIX2PIX	Yes	Tensorflow	0.0029 ± 0.0015	25.8669 ± 2.1632	0.8685 ± 0.0488	**0.0103**
Ours	Yes	Tensorflow	**0.0015 ± 0.0008**	**28.7602± 2.1338**	**0.9265 ± 0.0269**	0.0219

method $\widehat{I}_{smoke-free}$, as well as the estimated smoke mask $\widehat{I}_{mask}$. We found that most of the previous approaches can effectively remove a certain level of smoke, in which DCP seems to be the best one among the non-DL methods. But there are still many problems for these non-DL methods:

- Not robust enough to smoke variations (position, density, texture) and produced unstable results (eg. ATM)
- Cannot recover correct color for smoke-covered areas
- Color shift for non-smoke areas
- Suffer from over-saturated (eg. DCP, BCCR, DEFADE) or under-saturated (eg. ATM EVID, GMBI) problem

In contrast, our method can totally overcome these problems, which can not only focus on the smoke-covered areas and retain the smoke-free areas, but can also recover correct tissue colors based on the contextual knowledge learned by the network.

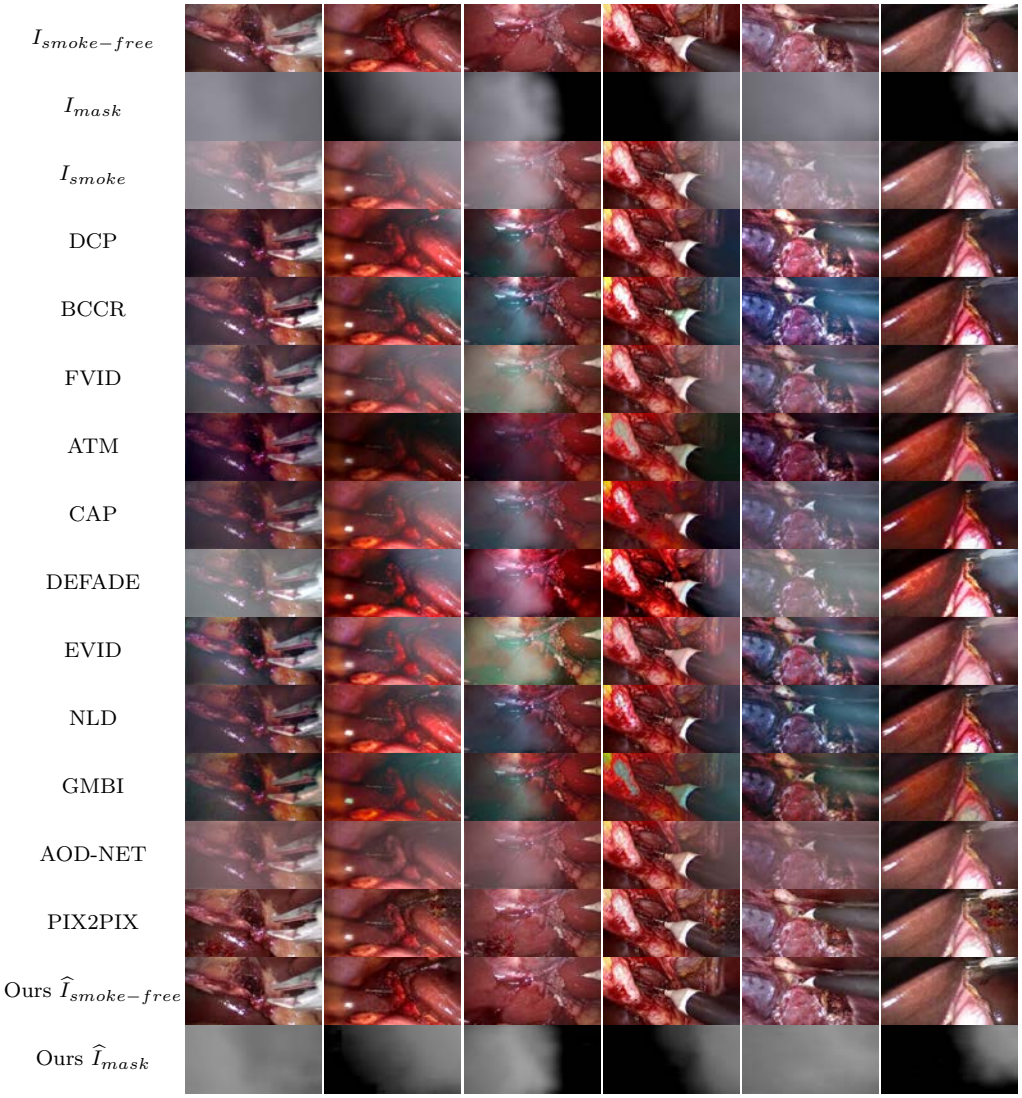


Figure 8.5: Qualitative results on synthetic testing dataset

AOD-NET's result is just above the conventional image processing based method as it uses a very shallow network structure. It is worth noticing that for GAN-based method like PIX2PIX, due to the characteristic of GAN loss, the network will learn to add some "fake" features to make the image looks like smoke-free image. However, these features are selected by the machine and totally uncontrollable. As can be seen from Fig. 8.5, the PIX2PIX network learned to add fake scars and reflections to the results, which is very harmful that can influence the surgeon's judgment if used during surgical interventions.

8.4.4 Smoke Removal Limit Test

Smoke can not only block structural information, but also fade color information. This loss of information is usually irreversible, depending on how thick the smoke is. To further evaluate the ability to recover smoke-free images under different smoke densities, we conduct a study of the de-smoking performance under 10 different smoke densities. We randomly selected 100 images from our 2005 test datasets and rendered 10 fixed-position smoke on each image with different smoke densities from 0 to 10, in which smoke level 0 means no rendered smoke, while smoke level 9 means the maximum smoke density.

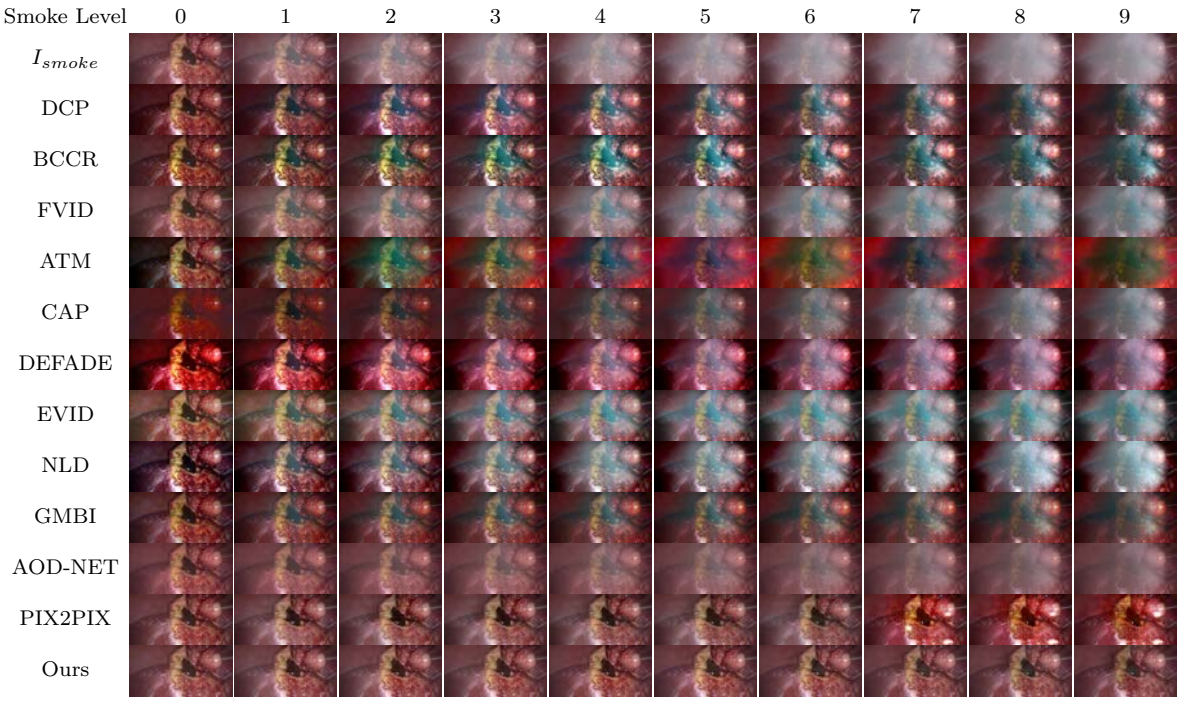


Figure 8.6: Qualitative result for our smoke removal density test.

As shown in Fig. 8.6, we present the rendered smoke images I_{smoke} in the first row with 10 smoke levels, and the de-smoked results from 11 previous methods, as well as our method. From the results, we found that most of the previous method cannot recover the correct color of the dark-red tissue in the center of the images. Also, a common problem of the previous method is that the estimated smoke-free images become blurry with the increase of smoke level. In contrast, DL-based methods can

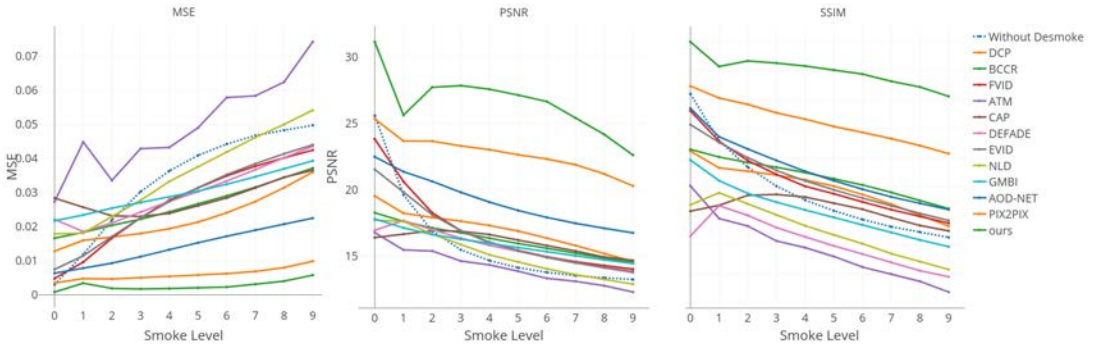


Figure 8.7: The quantitative results of our smoke removal limit test. From left to right: the MSE, PSNR and SSIM results for our method and 11 comparison approaches under 10 different smoke levels

give better results as the network learned to recover correct color based on contextual information. It is interesting to see that PIX2PIX produced similar results as ours, but became un-controllable after smoke level 7 that started to add "fake" reflections on the results. Our method produced very clean results with only minor saturation change, which is very hard to recover under very thick smoke.

For quantitative results in Fig. 8.7, we show the curves of the MSE, SSIM and PSNR between the image pairs of de-smoked image and smoke-free image for our result and 11 comparison methods under 10 different smoke levels. Our result yields the lowest MSE as well as highest SSIM and PSNR for all the 10 smoke levels, which significantly outperform all of the previous methods.

We also plotted the curves without any de-smoking process as a baseline. We found that the results for most of the approaches are worse than the baseline from the beginning even no smoke, but with the rise of the smoke level, the results become better than the baseline. This is because these approaches often result in the shift of color, increase in contrast and saturation, which would impact the error measurement over the first few smoke level. In contrast, our method shows very robust results to the rise of the smoke level, this is benefit from our novel learning frameworks that can recover correct tissue color under the circumstances of zero smoke and also very high smoke densities.

8.4.5 Evaluation on *in-vivo* data

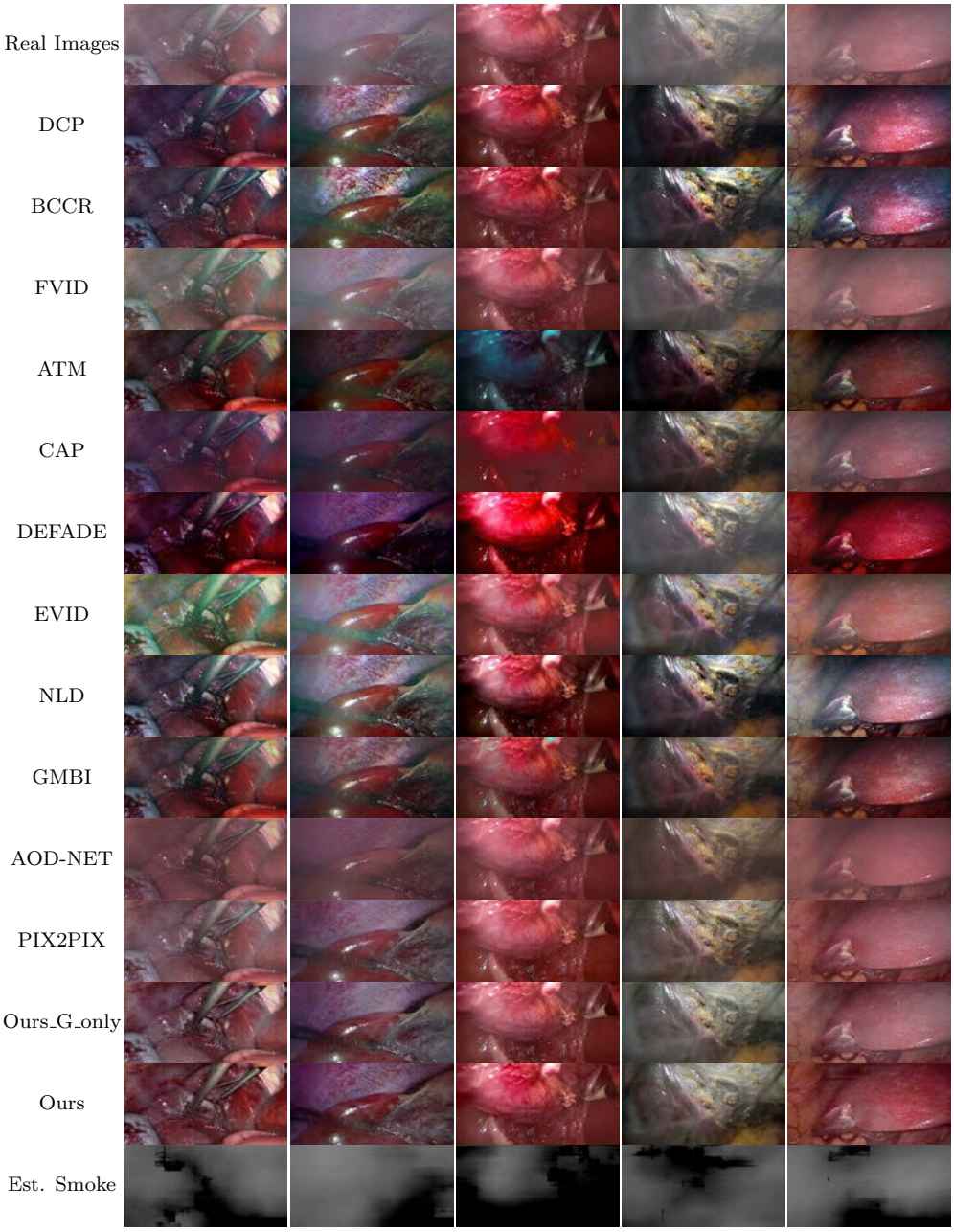


Figure 8.8: The qualitative results on *in-vivo* dataset.

Although our networks are trained purely on synthetic smoke images, we also evaluate our network on *in-vivo* data to test the ability to remove real surgical smoke. 81 images with the presence of smoke are manually picked from Hamlyn Centre Laparoscopic / Endoscopic Video Datasets and Cholec80 Dataset (Twinanda et al. 2017) for evaluation.

Fig. 8.8 shows the de-smoking visual results on *in-vivo* data. Again we found that some of the previous approaches either suffer from over-enhancement problem (such as DCP, BCCR, ATM, DEFADE) or cannot recover clear view (such as FVID, EVID). For DL-based methods, the color seems to be well recovered without over-enhacement. In details, we found that AOD-NET cannot recover clear view due to the use of very shallow network. For PIX2PIX, there is also some smoke remaining in the results. To fully understand the effectiveness of our GCN training framework, we also report the results of the generator-only version of our method as an ablation experiment. Our generator-only version gave similar results to PIX2PIX due to the similar network structure. When it comes to our results, we can see that all of the smoke is removed. The estimated smoke mask can correctly predict the real surgical smoke for the most of the time, but sometimes can fail such as the third one. The differences between our generator-only version and our final version prove that our smoke removal network is based on the predicted smoke mask, and the combination of smoke detection with smoke removal can narrow the gap between simulation and reality, improving the overall de-smoking performance for *in-vivo* dataset.

As there is no ground-truth smoke-free image pair from *in-vivo* data for quantitative evaluation, we adopt the Fog Aware Density Evaluator (FADE) (Choi et al. 2015) for the reference of perceptual smoke evaluation. FADE is a smoke prediction model based on natural scene statistics (NSS) and fog aware statistical features. The lower FADE score means the less perceptual fog and vice versa. The quantitative evaluation results by FADE are reported in Table 8.2, and we can find that our method didn't receive the lowest FADE score. This is because FADE is based on the statistics of the non-fog scene features, which will always take the image sharpness, contrast and saturation into consideration. However, our learning based method is trained and focused on recovering the natural and realistic smoke-free surgical images, but not emphasize the image visual quality metrics such as sharpness, contrast and saturation. For GAN-based method, from the previous experiments, we already know that it will create some fake features (such as scar) on the images to be looked like without smoke, so that PIX2PIX scores higher than our method. Fig. 8.8 shows

Table 8.2: FADE score on the *in-vivo* dataset from our method and 11 comparison approaches

Method	FADE Score	
	Avg.	Std.
DCP (He et al. 2011)	0.4315	0.1150
BCCR (Meng et al. 2013a)	0.3805	0.1147
FVAR (Galdran et al. 2016)	0.8722	0.2583
ATM (Sulami et al. 2014)	0.6582	1.7753
CAP (Zhu et al. 2015)	0.6082	0.2481
DEFADE (Choi et al. 2015)	0.6285	0.3993
EVAR (Galdran et al. 2015)	0.5383	0.1409
NLD (Berman et al. 2016)(Berman et al. 2017)	0.3693	0.1516
GMBI (Baid et al. 2017)	0.4259	0.0997
AOD-NET (Li et al. 2017a)	0.4871	0.1667
PIX2PIX (Isola et al. 2017)	0.4148	0.1044
Ours_G_Only	0.4647	0.1161
Ours	0.4465	0.1018

some example of de-smoking results on the *in-vivo* data

8.5 Discussion

8.5.1 Prevent Overfitting

One of the novelties of our work is that we do not require the ground truth data (the smoke and smoke-free image pairs) and can achieve unsupervised training from the view of data requirement. The *in-vivo* experiment proves that our networks although trained on synthetic data, can detect and remove smoke on real surgical data, which overcome the gap between simulation and reality. This is due to the fact that we adopt a lot of techniques to prevent our network overfitting to the synthetic data. For example, our training data is carefully selected and rendered: the backgrounds are extracted from 91 different laparoscopic and endoscopic videos

with different surgical procedures, different image color and tone, teh presence of different surgical instruments; the smoke is rendered by cinematic render engine with random intensities, densities, textures and positions. We believe that the decomposition of the de-smoking task into smoke detection and removal also helps to prevent overfitting. As we are not directly creating the mapping from smoke image to smoke-free image, but rather, we first detect the area and intensity of the smoke, then try to recover smoke-free image based on the smoke prior. We also intended to use a shallow network and drop-out for smoke detection to prevent overfitting. This solves the problem that deep learning need large amounts of hand-labelled ground truth for training, especially for medical datasets that professional knowledge is needed for labelling data.

8.5.2 Safety Issue

During the discussion with many medical practitioners, some concerns arose that removing the smoke from image might confuse the surgeons, as smoke although can block the view, but can also be a good signal for the on-going ablation process. These concerns inspired us to add the smoke detection network that can solve this problem by providing an extra pixel-level smoke detection before our smoke removal network remove the smoke. The predicted smoke can directly be shown to the surgeon or transferred to a different format for surgeons to receive it without distraction (see the potential application in next section).

It is also worth noticing that, although GAN framework (such as PIX2PIX) is a very good method for generating images, it can be very dangerous to be used for medical applications due to its uncertainty. During our experiments, we found that the GAN-based method can create fake "scars" or "reflections" to make the image looks like smoke-free image, which is totally unacceptable and may cause serious accidents if used during surgery.

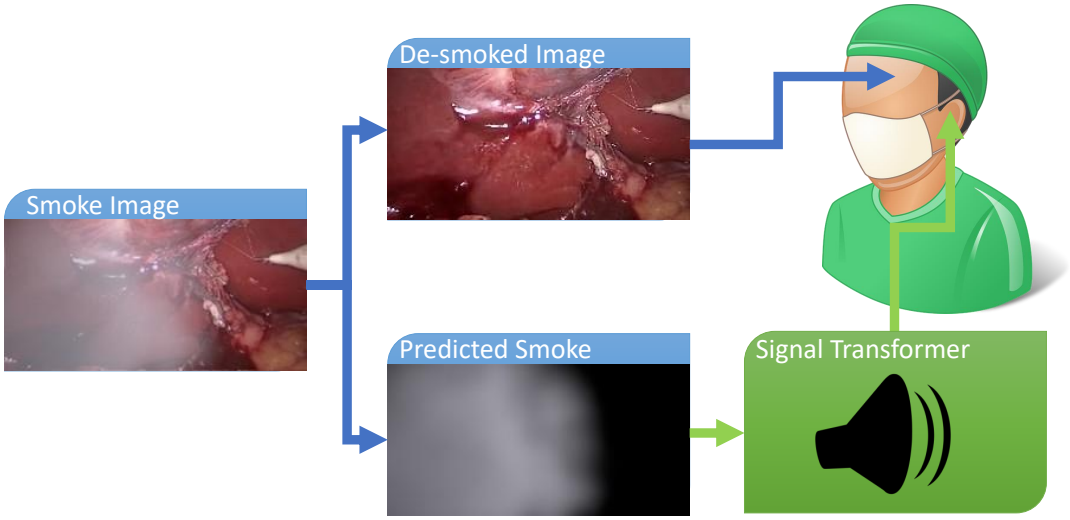
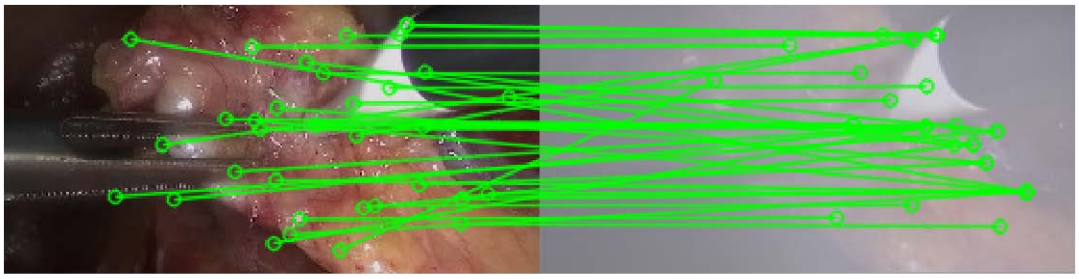


Figure 8.9: Potential application of our system: transforming smoke into sound

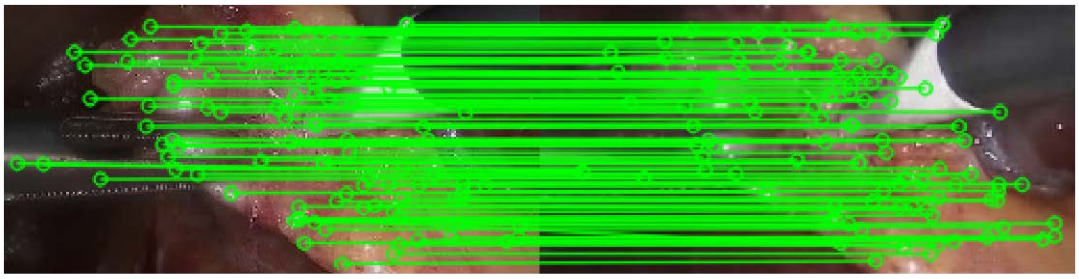
8.5.3 Application

Based on our smoke detection and removal framework, several advanced application can be built. One of it is related to the safety issue that we mentioned earlier that the surgical smoke is a good signal for the surgeon to know that the ablation is happening. As illustrated in Figure 8.9, our proposed method has the potential of transforming the predicted smoke into another format (such as sound) to alert the surgeons for the awareness of on-going ablation process, while watching the real-time de-smoked video stream.

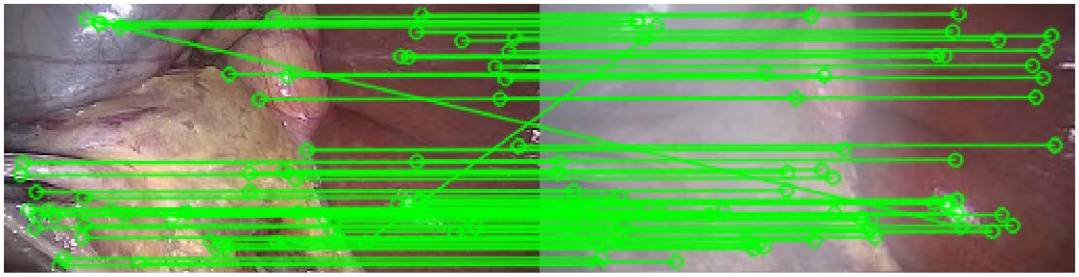
Also, the smoke removal is not only for surgeons but also can be used as a pre-processing step for many vision-based surgical assistance systems to improve the robustness to smoke such as tracking and reconstruction. As can be seen from Figure 8.10, we performed a standard SIFT (Scale Invariant Feature Transform) matching for the images before desmoking and after desmoking. The SIFT descriptors detected and matched are significantly increased after our desmoking framework. SIFT matching is an essential step for both tracking and reconstruction, the results shows that when using our desmoking framework as a pre-processing step, the tracking and reconstruction algorithms can be more robust and accurate.



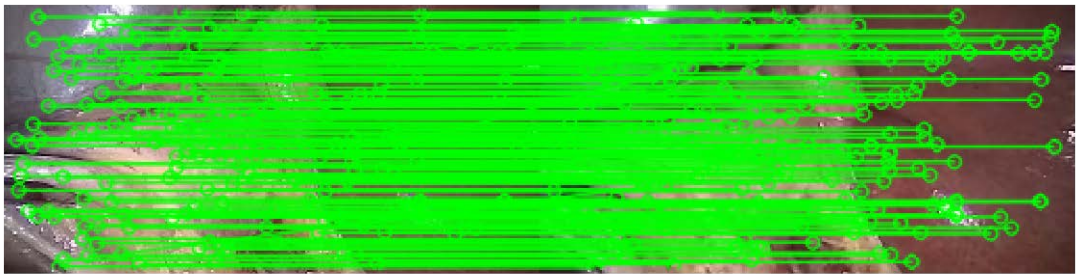
(a) Before desmoking – SIFT Detected: 57 / Matched: 35



(b) After desmoking – SIFT Detected: 391 / Matched: 93



(c) Before desmoking – SIFT Detected: 230 / Matched: 63



(d) After desmoking – SIFT Detected: 506 / Matched: 177

Figure 8.10: SIFT matching results. (a)(c) Before desmoking, (b)(d) after desmoking.

8.5.4 Future Work

In future work, we are going to combine CNNs with the recurrent neural networks (RNN) for video sequence smoke removal. Since during surgical ablation, the smoke density rise with time. RNN can help to memorize the features (such as tissue color) when there is light smoke and have the potential to recover them even with very high smoke density. It is also interesting to see whether training networks from synthetic dataset can be extended to many other tasks such as laparoscopic camera tracking, surgical instruments detection and tissue/organ segmentation, which will overcome the shortage of medical ground-truth data and greatly benefit the deep learning technology to be used in surgical scenes.

8.6 Conclusion

In this work, we present a novel deep learning framework for real-time surgical smoke detection and removal during minimally invasive surgery. Our unsupervised training framework only needs laparoscopic images as input, and 3D render engine is used to randomly render smoke on these images to synthesize datasets for training. The novelty of this work lies in our GCN training framework that used the smoke detection network as prior knowledge and also a supervision for our smoke removal network. With this initiative, We not only achieve pixel-level smoke detection, but also help improve the smoke removal performance compared to the state-of-the-art smoke removal methods. Our framework also yields extra benefit of preventing over-fitting to synthetic datasets, and also have many potential applications for surgical human-computer interactions.

Chapter 9

Conclusions and Future Work

This thesis has presented three pieces of work on online surface reconstruction in AR for the perception of the surrounding environment, as well as two pieces of work for high-level contextual understanding and increasing robustness of the reconstruction. Our proposed work has several key clinical and real-world applications such as surgical AR guidance, intra-body measurement, as well as AR games with realistic virtual-real interactions when using our novel context-aware AR. Additionally, with our proposed learning-based desmoking networks, our tracking and reconstruction can work more robust under extremely surgical environment.

9.1 Achievements of This Thesis

The main achievements of this thesis are:

In Chapter 2, a novel literature classification method by combining text mining, topic generation/clustering, and taxonomic review is presented for a better understanding of development trends, current issues and future directions. This provides an efficient method for researchers when finding papers and identifying topics in a new research area, providing an automatic classification and generation of potential research topics. Based on the classification results, the trends of different research topics can be easily accessed, which gives the researchers hints for current research trends and future developments. Our classification led to SLAM, Reconstruction

and Deep Learning for AR being the research focus for this thesis.

In Chapter 4, a SLAM-based monocular camera tracking and dense reconstruction method is proposed for geometry-aware AR in a challenging MIS environment. A series of surface reconstruction technologies are employed to compile a global dense surface from sparse landmark points. The evaluation experiments show promising reconstruction results. However, this method relies on consecutive frames for reconstruction, which is sensitive to deformation, blood, surgical smoke and the movement of surgical instruments.

To solve these problems, in Chapter 5, a novel stereo-based on-the-fly reconstruction framework is proposed. With the theory of stereo-vision epipolar constraint, the depth can be accurately estimated for a single frame pair. Followed by the back-end SLAM system, the global reconstruction can be incrementally built. This method is more robust to noise and deformation, but requires stereo endoscopic cameras, which are not widely used.

With the recent success of Convolutional Neural Networks (CNNs), many ill-posed problems, such as single image depth prediction are not impossible anymore. In Chapter 6, we present a novel learning-based single image depth estimation with confidence for accurate and reliable single image 3D reconstruction. Evaluation on public datasets shows that our method outperforms the state-of-the-art results.

In Chapter 7, we are not content with the uniform geometry-aware AR and introduce a contextual understanding of the environment with the help of deep learning. An interactive context-aware AR framework is proposed based on the latest SLAM technology and learning based material recognition for providing a whole new AR experience. An accuracy experiment and user study show that our method can accurately deliver context-based interaction and greatly increased the AR experience compared with the geometry-aware AR.

In Chapter 8, a learning-based smoke removal approach is proposed for increasing the tracking and reconstruction robustness in the challenging MIS environment with the absence of surgical smoke. A novel generative-collaborative learning scheme is presented that decomposes the de-smoke task into two separate tasks: smoke

detection and smoke removal. While using the detection network as prior knowledge, it is also used as a loss function to maximize its support for the removal of network training. The quantitative and qualitative studies show that our training framework outperforms the latest GAN framework (such as PIX2PIX) and the state-of-the-art de-smoking approaches.

9.2 Conclusions

In this thesis, a series of methods are proposed to solve the challenging dense 3D surface reconstruction problem. To conclude:

- Dense 3D surface can be reconstructed from monocular camera sequences using the sparse point cloud from SLAM system. (Chapter 3)
- Global Dense 3D surface can be incrementally reconstructed using stereo matching combined with SLAM system. (Chapter 4)
- Dense 3D surface can be reconstructed from single monocular image using learning based depth prediction. (Chapter 5)
- Dense semantic 3D surface can be reconstructed using deep semantic scene understanding methodology combined with dense 3D scene reconstruction, which can build high-level context-aware highly interactive MR environment. (Chapter 6)
- Under extreme conditions such as the presence of surgical smoke, our proposed learning based smoke removal can recover a clear view for accurate and robust 3D surface reconstruction. (Chapter 7)

9.3 Discussions and Future Perspectives

Although our proposed tracking and reconstruction methods have been carefully evaluated on both synthetic datasets and *in-vivo* datasets with promising results

and applications, it is still far away for these technologies to be used in real-world scenarios and medical practices, due to the following reasons:

- Accuracy is a critical property that directly determines whether or not an AR system could be used in medical practice. If the AR content were superimposed in the wrong position, then the surgeon would be misled into making a wrong decision, which could cause a serious medical accident. Real time performance, and minimum latency are pre-requisites in most medical applications and directly affect the usability of AR. Based on the research of (Lambert et al. 2016), it should be as high as 100 FPS at least to enable the doctor to detect minor changes in a surgical video. However, most of AR systems are very complex and can only be running at 20FPS.

- Robustness is the entry criteria of AR to be widely used in medical and real-world scenarios. Especially in surgical scenes that tissue movement, surgical smoke, occlusion of instruments can make the tracking and reconstruction in-accurate and even fail. This is because SLAM theory is developed based on static world assumption; the deformations of objects (such as tissues and organs) directly challenge this basic condition for SLAM to estimate camera poses for 3D reconstruction. Therefore, soft tissue deformation is a great challenge to support in the SLAM based reconstruction framework as proposed here. Particularly with monocular endoscopic videos, it is extremely hard to recover the soft deformation correctly while simultaneously estimating the camera poses. For small deformations like those in the *in-vivo* video that we use, however, the RANSAC algorithm in SLAM system will filter the outliers and recover the correct movement. For large deformation in very small FOVs, it is still unclear how to solve the tissue deformation issue without using extra external sensors within the monocular scene.

Although the road is very tough, the future is promising. It is totally foreseeable that in the near future, machine and AI will be a great assistance for making human's life and work easier by either providing additional information (eg. surgical

guidance) or autonomously take over some repetitive job (eg. self-driving car). It is no doubt that AR will be an important bridge for connecting machine with human, and become a very essential human-computer interface that replaces screens that have a border. Therefore, sensing and perceiving the world's geometry and context are the very first and enabling technologies for AR's promising future, which is the meaning of this PhD project.

For the future development of AR, the hardware evolution is very important. Current AR headsets such as HoloLens is very expensive, bulky, heavy, slow and battery life is short, which limit the AR entering everyone's life. Also, better hardware can make the result more accurate and the algorithm faster, easier and more robust. A good example learnt from this PhD project is that for imaging sensor, monocular camera needs complex SLAM system or machine learning for reconstruction. But with stereo camera, more accurate results can be computed from easier stereo matching algorithm; with depth camera such as Kinect, the reconstruction results are much more accurate and reliable. However, in MIS, the most common available sensor is still monocular camera, which to some extent limits the development of AR in MIS.

As the next generation of human-machine interface, the intelligent interaction is also a very interesting research topic. In Chapter 6, we initiatively integrate machine learning into AR framework for context-aware applications. The results prove that with the power of AI, AR can be more natural and immersive, leading to better user experience. However, this work has addressed a specialised application and only started to explore the use of a machine learning model. It can be predicted that AI will be an important part of AR for more general, persistent and intelligent applications.

References

Abe, Y., Sato, S., Kato, K., Hyakumachi, T., Yanagibashi, Y., Ito, M. and Abumi, K., 2013. A novel 3D guidance system using augmented reality for percutaneous vertebroplasty: technical note. *Journal of Neurosurgery: Spine*, 19 (4), 492–501.

Agrawal, M., Konolige, K. and Blas, M. R., 2008. Censure: Center surround extremas for realtime feature detection and matching. *Computer Vision–ECCV 2008*, Springer, 102–115.

Alahi, A., Ortiz, R. and Vandergheynst, P., 2012. Freak: Fast retina keypoint. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, Ieee, 510–517.

Alamri, A., Cha, J. and El Saddik, A., 2010. Ar-rehab: An augmented reality framework for poststroke-patient rehabilitation. *Instrumentation and Measurement, IEEE Transactions on*, 59 (10), 2554–2563.

Albrecht, U.-V., von Jan, U., Kuebler, J., Zoeller, C., Lacher, M., Muensterer, O. J., Ettinger, M., Klintschar, M. and Hagemeier, L., 2013. Google glass for documentation of medical findings: evaluation in forensic medicine. *Journal of medical Internet research*, 16 (2), e53–e53.

Alemzadeh, H., Raman, J., Leveson, N., Kalbarczyk, Z. and Iyer, R. K., 2016. Adverse events in robotic surgery: A retrospective study of 14 years of fda data. *PLOS ONE*, 11 (4), e0151470.

Andersen, D., Popescu, V., Cabrera, M. E., Shanghavi, A., Gpmez, G., Marley, S.,

Mullis, B. and Wachs, J., 2016. Avoiding focus shifts in surgical telementoring using an augmented reality transparent display. *MMVR 22*, 9–14.

Anderson, F., Grossman, T., Matejka, J. and Fitzmaurice, G., 2013. Youmove: enhancing movement training with an augmented reality mirror. *Proceedings of the 26th annual ACM symposium on User interface software and technology*, ACM, 311–320.

Apple, 2017. Arkit - apple developer. URL `https://developer.apple.com/arkit/`.

Apple, 2018. metaio. https://www.metaio.com/. [Accessed 29 4 2016].

Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M. and Savarese, S., 2016. 3d semantic parsing of large-scale indoor spaces. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1534–1543. URL `http://dx.doi.org/10.1109/CVPR.2016.170`.

Assad-Kottner, C., Hakeem, A., Fontenot, E. and Uretsky, B. F., 2014. "telementoring": an interventional procedure using a wearable computer: first-in-man. *Journal of the American College of Cardiology*, 63 (10), 1022–1022.

Azimi, E., Doswell, J. and Kazanzides, P., 2012. Augmented reality goggles with an integrated tracking system for navigation in neurosurgery. *Virtual Reality Short Papers and Posters (VRW), 2012 IEEE*, IEEE, 123–124.

Azuma, R. T., 1997. A survey of augmented reality. *Presence: Teleoperators and virtual environments*, 6 (4), 355–385.

Badrinarayanan, V., Kendall, A. and Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for scene segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP (99), 1–1.

Baid, A., Kotwal, A., Bhalodia, R., Merchant, S. N. and Awate, S. P., 2017. Joint desmoking, specularity removal, and denoising of laparoscopy images via graphical models and bayesian inference. *IEEE ISBI*.

Bailey, T., Nieto, J., Guivant, J., Stevens, M. and Nebot, E., 2006. Consistency of the EKF-SLAM algorithm. *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, IEEE, 3562–3568.

Bansal, R., Raj, G. and Choudhury, T., 2016. Blur image detection using laplacian operator and open-CV. *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*, IEEE.

Barnard, S. T. and Fischler, M. A., 1982. Computational stereo. *ACM Comput. Surv.*, 14 (4), 553–572. URL `http://doi.acm.org/10.1145/356893.356896`.

Barreira, J., Bessa, M., Barbosa, L. and Magalh'es, L., 2018. A context-aware method for authentically simulating outdoors shadows for mobile augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 24 (3), 1223–1231.

Bay, H., Tuytelaars, T. and Van Gool, L., 2006. Surf: Speeded up robust features. *Computer vision–ECCV 2006*, Springer, 404–417.

Bell, S., Upchurch, P., Snavely, N. and Bala, K., 2015. Material recognition in the wild with the materials in context database. *Computer Vision and Pattern Recognition (CVPR)*.

Berman, D., Treibitz, T. and Avidan, S., 2016. Non-local image dehazing. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE.

Berman, D., Treibitz, T. and Avidan, S., 2017. Air-light estimation using haze-lines. *2017 IEEE International Conference on Computational Photography (ICCP)*, IEEE.

Bernhardt, S., Nicolau, S. A., Agnus, V., Soler, L., Doignon, C. and Marescaux, J., 2016. Automatic localization of endoscope in intraoperative CT image: A simple approach to augmented reality guidance in laparoscopic surgery. *Medical image analysis*, 30, 130–143.

Best, P. J. and McKay, N. D., 1992. A method for registration of 3-D shapes. *IEEE Transactions on pattern analysis and machine intelligence*, 14 (2), 239–256.

Blei, D. M., Ng, A. Y. and Jordan, M. I., 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3, 993–1022.

Blender, 2016. Blender - free and open 3d creation software. URL `https://www.blender.org/`, [Accessed 6 Nov. 2016].

Blum, T., Kleeberger, V., Bichlmeier, C. and Navab, N., 2012a. mirracle: An augmented reality magic mirror system for anatomy education. *Virtual Reality Short Papers and Posters (VRW), 2012 IEEE*, IEEE, 115–116.

Blum, T., Stauder, R., Euler, E. and Navab, N., 2012b. Superman-like x-ray vision: towards brain-computer interfaces for medical augmented reality. *Mixed and Augmented Reality (ISMAR), 2012 IEEE International Symposium on*, IEEE, 271–272.

Bolkar, S., Wang, C. and nad Sule Yildirim, F. A. C., 2018. Deep smoke rremoval from mimimally invasive surgery videos. *2018 IEEE International Conference on Image Processing (ICIP)*.

Boonbrahm, P. and Kaewrat, C., 2014. Assembly of the virtual model with real hands using augmented reality technology. *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments*, Springer, 329–338.

Bourdel, N., Collins, T., Pizarro, D., Debize, C., sophie Grémeau, A., Bartoli, A. and Canis, M., 2017. Use of augmented reality in laparoscopic gynecology to visualize myomas. *Fertility and Sterility*.

Bretón-López, J., Quero, S., Botella, C., García-Palacios, A., Baños, R. M. and Alcañiz, M., 2010. An augmented reality system validation for the treatment of cockroach phobia. *Cyberpsychology, Behavior, and Social Networking*, 13 (6), 705–710.

Cai, B., Xu, X., Jia, K., Qing, C. and Tao, D., 2016. DehazeNet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25 (11), 5187–5198.

Calonder, M., Lepetit, V., Strecha, C. and Fua, P., 2010. Brief: Binary robust independent elementary features. *Computer Vision–ECCV 2010*, 778–792.

Calvet, L., Gurdjos, P. and Charvillat, V., 2012. Camera tracking using concentric circle markers: paradigms and algorithms. *Image Processing (ICIP), 2012 19th IEEE International Conference on*, IEEE, 1361–1364.

Cao, Y., Wu, Z. and Shen, C., 2017. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, PP (99), 1.

Carmigniani, J., Furht, B., Anisetti, M., Ceravolo, P., Damiani, E. and Ivkovic, M., 2011. Augmented reality technologies, systems and applications. *Multimedia Tools and Applications*, 51 (1), 341–377.

Castle, R., Klein, G. and Murray, D. W., 2008. Video-rate localization in multiple maps for wearable augmented reality. *Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on*, IEEE, 15–22.

Chae, Y. S., Lee, S. H., Lee, H. K. and Kim, M. Y., 2015. Optical coordinate tracking system using afocal optics for image-guided surgery. *International journal of computer assisted radiology and surgery*, 10 (2), 231–241.

Chang, P.-L., Handa, A., Davison, A. J., Stoyanov, D. and Edwards, P. E., 2014a. Robust real-time visual odometry for stereo endoscopy using dense quadrifocal tracking. *Information Processing in Computer-Assisted Interventions*, Springer Science + Business Media, 11–20.

Chang, P.-L., Handa, A., Davison, A. J., Stoyanov, D. et al., 2014b. Robust real-time visual odometry for stereo endoscopy using dense quadrifocal tracking. *Interna-*

tional Conference on Information Processing in Computer-Assisted Interventions, Springer, 11–20.

Chang, P.-L., Stoyanov, D., Davison, A. J. and Edwards, P. E., 2013. Real-time dense stereo reconstruction using convex optimisation with a cost-volume for image-guided robotic surgery. *Med Image Comput Comput Assist Interv*, 16 (Pt 1), 42–49.

Chang, T.-C., Hsieh, C.-H., Huang, C.-H., Yang, J.-W., Lee, S.-T., Wu, C.-T. and Lee, J.-D., 2015. Interactive medical augmented reality system for remote surgical assistance. *Appl. Math*, 9 (1L), 97–104.

Chen, L., , Tang, W. and John, N. W., 2018a. Unsupervised learning of surgical smoke removal from simulation. *Hamlyn Symposium on Medical Robotics*.

Chen, L., 2016. Supplemental video. URL `https://youtu.be/Y9D3Liw5tXo`, [Accessed 12 Feb. 2011].

Chen, L., Day, T. W., Tang, W. and John, N. W., 2017a. Recent developments and future challenges in medical mixed reality. *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, IEEE.

Chen, L., Francis, K. and Tang, W., 2017b. Semantic augmented reality environment with material-aware physical interactions. *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, IEEE.

Chen, L., Tang, W. and John, N. W., 2017c. Real-time geometry-aware augmented reality in minimally invasive surgery. *Healthcare Technology Letters*, 4 (5), 163–167.

Chen, L., Tang, W., John, N. W., Wan, T. R. and Zhang, J. J., 2018b. SLAM-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality. *Computer Methods and Programs in Biomedicine*, 158, 135–146.

Chen, W., Fu, Z., Yang, D. and Deng, J., 2016. Single-image depth perception in the wild. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett, eds., *Advances in Neural Information Processing Systems 29*, Curran Associates, Inc., 730–738. URL `http://papers.nips.cc/paper/6489-single-image-depth-perception-in-the-wild.pdf`.

Chen, X., Wang, L., Fallavollita, P. and Navab, N., 2013. Precise x-ray and video overlay for augmented reality fluoroscopy. *International journal of computer assisted radiology and surgery*, 8 (1), 29–38.

Chen, Y.-S., Hung, Y.-P. and Fuh, C.-S., 2001. Fast block matching algorithm based on the winner-update strategy. *IEEE Transactions on Image Processing*, 10 (8), 1212–1222.

Chevaillier, P., Trinh, T.-H., Barange, M., De Loor, P., Devillers, F., Soler, J. and Querrec, R., 2012. Semantic modeling of virtual environments using mascaret. *Software Engineering and Architectures for Realtime Interactive Systems (SEARIS), 2012 5th Workshop on*, IEEE, 1–8.

Choi, L. K., You, J. and Bovik, A. C., 2015. Referenceless prediction of perceptual fog density and perceptual image defogging. *IEEE Transactions on Image Processing*, 24 (11), 3888–3901.

Choi, Y., 2011. Ubi-rehab: An android-based portable augmented reality stroke rehabilitation system using the eglove for multiple participants. *Virtual Rehabilitation (ICVR), 2011 International Conference on*, IEEE, 1–2.

Civera, J., Grasa, O. G., Davison, A. J. and Montiel, J., 2010. 1-point ransac for extended Kalman filtering: Application to real-time structure from motion and visual odometry. *Journal of Field Robotics*, 27 (5), 609–631.

Coles, T., Meglan, D. and John, N. W., 2011a. The role of haptics in medical training simulators: a survey of the state of the art. *Haptics, IEEE Transactions on*, 4 (1), 51–66.

Coles, T. R., John, N. W., Gould, D. A. and Caldwell, D. G., 2011b. Integrating haptics with augmented reality in a femoral palpation and needle insertion training simulation. *Haptics, IEEE Transactions on*, 4 (3), 199–209.

Cosentino, F., John, N. W. and Vaarkamp, J., 2017. Rad-ar: Radiotherapy - augmented reality. *2017 International Conference on Cyberworlds (CW)*, 226–228.

Daly, M., Chan, H., Prisman, E., Vescan, A., Nithiananthan, S., Qiu, J., Weersink, R., Irish, J. and Siewerdsen, J., 2010. Fusion of intraoperative cone-beam ct and endoscopic video for image-guided procedures. *SPIE Medical Imaging*, International Society for Optics and Photonics, 762503–762503.

DAQRI, 2016. Artoolkit.org: Open source augmented reality sdk. http://artoolkit.org/. [Accessed 29 4 2016].

Davison, A. J., Reid, I. D., Molton, N. D. and Stasse, O., 2007. Monoslam: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29 (6), 1052–1067.

De Marsico, M., Levialdi, S., Nappi, M. and Ricciardi, S., 2014. Figi: floating interface for gesture-based interaction. *Journal of Ambient Intelligence and Humanized Computing*, 5 (4), 511–524.

De Paolis, L. T. and Aloisio, G., 2010. Augmented reality in minimally invasive surgery. *Advances in Biomedical Sensing, Measurements, Instrumentation and Systems*, Springer, 305–320.

De Troyer, O., Kleinermann, F., Pellens, B. and Bille, W., 2007. Conceptual modeling for virtual reality. *Tutorials, Posters, Panels and Industrial Contributions at the 26th International Conference on Conceptual Modeling - Volume 83*, Darlinghurst, Australia, Australia: Australian Computer Society, Inc., ER '07, 3–18. URL `http://dl.acm.org/citation.cfm?id=1386957.1386959`.

Debarba, H. G., Grandi, J., Maciel, A. and Zanchet, D., 2012. Anatomic hepatectomy planning through mobile display visualization and interaction. *MMVR*, 111–115.

Deng, H., Zhang, L., Mao, X. and Qu, H., 2016. Interactive urban context-aware visualization via multiple disocclusion operators. *IEEE Transactions on Visualization and Computer Graphics*, 22 (7), 1862–1874.

Dey, A., Billinghurst, M., Lindeman, R. W. and Swan II, J. E., 2016. A systematic review of usability studies in augmented reality between 2005 and 2014. *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*.

Dissanayake, M., Newman, P., Clark, S., Durrant-Whyte, H. F. and Csorba, M., 2001. A solution to the simultaneous localization and map building (slam) problem. *Robotics and Automation, IEEE Transactions on*, 17 (3), 229–241.

Dosovitskiy, A., Springenberg, J. T., Tatarchenko, M. and Brox, T., 2017. Learning to generate chairs, tables and cars with convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (4), 692–705.

Du, X., Clancy, N., Arya, S., Hanna, G. B., Kelly, J., Elson, D. S. and Stoyanov, D., 2015. Robust surface tracking combining features, intensity and illumination compensation. *International journal of computer assisted radiology and surgery*, 10 (12), 1915–1926.

Dunkin, B. J. and Flowers, C., 2015. 3d in the minimally invasive surgery (mis) operating room: Cameras and displays in the evolution of mis. *Imaging and Visualization in The Modern Operating Room*, Springer, 145–155.

Eck, U., Pankratz, F., Sandor, C., Klinker, G. and Laga, H., 2014. Comprehensive workspace calibration for visuo-haptic augmented reality. *Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on*, IEEE, 123–128.

Edgcumbe, P., Pratt, P., Yang, G.-Z., Nguan, C. and Rohling, R., 2015. Pico lantern:

Surface reconstruction and augmented reality in laparoscopic surgery using a pick-up laser projector. *Medical image analysis*, 25 (1), 95–102.

Egui Zhu, N. Z., 2014. Pedagogy of mobile augmented reality in health education. *Proc. Int Interactive Mobile Communication Technologies and Learning (IMCL) Conf*, 209–212.

Eigen, D. and Fergus, R., 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *2015 IEEE International Conference on Computer Vision (ICCV)*, 2650–2658.

Eigen, D., Puhrsch, C. and Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, Cambridge, MA, USA: MIT Press, NIPS'14, 2366–2374. URL `http://dl.acm.org/citation.cfm?id=2969033.2969091`.

Elhawary, H. and Popovic, A., 2011. Robust feature tracking on the beating heart for a robotic-guided endoscope. *The international journal of medical robotics and computer assisted surgery*, 7 (4), 459–468.

Elsevier, 2018. Content - scopus — elsevier. https://www.elsevier.com/solutions/scopus/content.

Endert, A., Fiaux, P. and North, C., 2012. Semantic interaction for visual text analytics. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, CHI '12, 473–482. URL `http://doi.acm.org/10.1145/2207676.2207741`.

Engel, J., Schöps, T. and Cremers, D., 2014. LSD-SLAM: Large-scale direct monocular SLAM. *Computer Vision – ECCV 2014*, Springer Nature, 834–849.

Erazo, O., Pino, J. A., Pino, R. and Fernandez, C., 2014. Magic mirror for neurorehabilitation of people with upper limb dysfunction using kinect. *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, IEEE, 2607–2615.

Fattal, R., 2008. Single image dehazing. *ACM Transactions on Graphics*, 27 (3), 1.

Fiala, M., 2005. Artag, a fiducial marker system using digital techniques. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, IEEE, volume 2, 590–596.

Fischbach, M., Wiebusch, D. and Latoschik, M. E., 2017. Semantic entity-component state management techniques to enhance software quality for multimodal vr-systems. *IEEE Transactions on Visualization and Computer Graphics*, 23 (4), 1342–1351.

Fischler, M. A. and Bolles, R. C., 1981a. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6), 381–395.

Fischler, M. A. and Bolles, R. C., 1981b. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24 (6), 381–395. URL `http://doi.acm.org/10.1145/358669.358692`.

Fitzgibbon, A., Wexler, Y. and Zisserman, A., 2003. Image-based rendering using image-based priors. *2003 IEEE International Conference on Computer Vision (ICCV)*, 1176–1183 vol.2.

Flynn, J., Neulander, I., Philbin, J. and Snavely, N., 2016. Deep stereo: Learning to predict new views from the world's imagery. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5515–5524.

Franz, A. M., Haidegger, T., Birkfellner, W., Cleary, K., Peters, T. M. and Maier-Hein, L., 2014. Electromagnetic tracking in medicine - a review of technology, validation, and applications. *Medical Imaging, IEEE Transactions on*, 33 (8), 1702–1725.

Fritz, J., U-Thainual, P., Ungi, T., Flammang, A. J., Cho, N. B., Fichtinger, G., Iordachita, I. I. and Carrino, J. A., 2012. Augmented reality visualization with

image overlay for mri-guided intervention: accuracy for lumbar spinal procedures with a 1.5-t mri system. *American Journal of Roentgenology*, 198 (3), W266–W273.

Fuchs, H., Livingston, M. A., Raskar, R., Colucci, D., Keller, K., State, A., Crawford, J. R., Rademacher, P., Drake, S. H. and Meyer, A. A., 1998. Augmented reality visualization for laparoscopic surgery. W. M. Wells, A. Colchester and S. Delp, eds., *Medical Image Computing and Computer-Assisted Intervention - Miccai'98*, Springer Berlin Heidelberg, volume 1496, 934–943.

Furness III, T. A. and Kollin, J. S., 1995. Virtual retinal display.

Galdran, A., Vazquez-Corral, J., Pardo, D. and Bertalmío, M., 2015. Enhanced variational image dehazing. *SIAM Journal on Imaging Sciences*, 8 (3), 1519–1546.

Galdran, A., Vazquez-Corral, J., Pardo, D. and Bertalmio, M., 2016. Fusion-based variational image dehazing. *IEEE Signal Processing Letters*, 1–1.

Garcia, J. A. and Navarro, K. F., 2014. The mobile rehapp: an ar-based mobile game for ankle sprain rehabilitation. *Serious Games and Applications for Health (SeGAH), 2014 IEEE 3rd International Conference on*, IEEE, 1–6.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V. and Garcia-Rodriguez, J., 2017. A review on deep learning techniques applied to semantic segmentation.

Garg, R., B.G., V. K., Carneiro, G. and Reid, I., 2016. Unsupervised cnn for single view depth estimation: Geometry to the rescue. B. Leibe, J. Matas, N. Sebe and M. Welling, eds., *Computer Vision – ECCV 2016*, Cham: Springer International Publishing, 740–756.

Gatrell, L. B., Hoff, W. A. and Sklair, C. W., 1992. Robust image features: Concentric contrasting circles and their image extraction. *Robotics-DL tentative*, International Society for Optics and Photonics, 235–244.

Gireesh, A. G., Gowda, M. et al., 2008. Acm transactions on information systems (1989–2006): A bibliometric study. *Information Studies*, 14 (4), 223–234.

Godard, C., Aodha, O. M. and Brostow, G. J., 2017. Unsupervised monocular depth estimation with left-right consistency. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6602–6611.

Google, 2017. Arcore - google developers. URL `https://developers.google.com/ar/`.

Google, 2018. Google glass. https://www.google.com/glass/start/. [Accessed 29 4 2016].

Grandi, J., Maciel, A., Debarba, H. and Zanchet, D., 2014. Spatially aware mobile interface for 3d visualization and interactive surgery planning. *Serious Games and Applications for Health (SeGAH), 2014 IEEE 3rd International Conference on*, IEEE, 1–8.

Grasa, O. G., Bernal, E., Casado, S., Gil, I. and Montiel, J., 2014. Visual slam for handheld monocular endoscope. *Medical Imaging, IEEE Transactions on*, 33 (1), 135–146.

Grasa, O. G., Civera, J., Guemes, A., Munoz, V. and Montiel, J., 2009. Ekf monocular slam 3d modeling, measuring and augmented reality from endoscope image sequences. *Medical image computing and computer-assisted intervention (MICCAI)*, volume 2.

Grasa, O. G., Civera, J. and Montiel, J., 2011. Ekf monocular slam with relocalization for laparoscopic sequences. *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, IEEE, 4816–4821.

Grubert, J., Langlotz, T., Zollmann, S. and Regenbrecht, H., 2017. Towards pervasive augmented reality: Context-awareness in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 23 (6), 1706–1724.

Hanna, M. G., Ahmed, I., Nine, J., Prajapati, S. and Pantanowitz, L., 2018. Augmented reality technology using microsoft hololens in anatomic pathology. *Archives of Pathology & Laboratory Medicine*, 142 (5), 638–644. PMID: 29384690.

Haouchine, N., Cotin, S., Peterlik, I., Dequidt, J., Lopez, M. S., Kerrien, E. and Berger, M.-O., 2015. Impact of soft tissue heterogeneity on augmented reality for liver surgery. *Visualization and Computer Graphics, IEEE Transactions on*, 21 (5), 584–597.

Haouchine, N., Dequidt, J., Peterlik, I., Kerrien, E., Berger, M.-O. and Cotin, S., 2013. Image-guided simulation of heterogeneous tissue deformation for augmented reality during hepatic surgery. *Mixed and Augmented Reality (ISMAR), 2013 IEEE International Symposium on*, IEEE, 199–208.

Haouchine, N., Dequidt, J., Peterlik, I., Kerrien, E., Berger, M.-O. and Cotin, S., 2014. Towards an accurate tracking of liver tumors for augmented reality in robotic assisted surgery. *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, IEEE, 4121–4126.

He, K., Sun, J. and Tang, X., 2011. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33 (12), 2341–2353.

Hempel, J., 2015. Microsoft in the age of satya nadella. URL `http://www.wired.com/2015/01/microsoft-nadella/`, [Accessed 2 5 2016].

Hermans, A., Floros, G. and Leibe, B., 2014. Dense 3d semantic mapping of indoor scenes from rgb-d images. *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, IEEE, 2631–2638.

Hirschmuller, H., 2008. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30 (2), 328–341.

Hochreiter, J., Daher, S., Nagendran, A., Gonzalez, L. and Welch, G., 2015. Touch sensing on non-parametric rear-projection surfaces: A physical-virtual head for hands-on healthcare training. *Virtual Reality (VR), 2015 IEEE*, IEEE, 69–74.

Höllerer, T. and Feiner, S., 2004. Mobile augmented reality. *Telegeoinformatics: Location-Based Computing and Services. Taylor and Francis Books Ltd., London, UK*, 21.

Holmdahl, T., 2015. Build 2015: A closer look at the microsoft hololens hardware. https://blogs.windows.com/devices/2015/04/30/build-2015-a-closer-look-at-the-microsoft-hololens-hardware/. [Accessed 29 4 2016].

Horn, B. K. and Schunck, B. G., 1981. Determining optical flow. *1981 Technical symposium east*, International Society for Optics and Photonics, 319–331.

HOSTETTLER, A., FOREST, C., SOLER, L. and MARESCAUX, J., 2011. A cost effective simulator for education of ultrasound image interpretation and probe manipulation. *Medicine Meets Virtual Reality 18: NextMed*, 163, 403.

Hsieh, C.-H. and Lee, J.-D., 2015. Markerless augmented reality via stereo video see-through head-mounted display device. *Mathematical Problems in Engineering*, 2015.

Hu, L., Wang, M. and Song, Z., 2013. A convenient method of video see-through augmented reality based on image-guided surgery system. *Internet Computing for Engineering and Science (ICICSE), 2013 Seventh International Conference on*, IEEE, 100–103.

Hua, Y. and Tian, H., 2016. Depth estimation with convolutional conditional random field network. *Neurocomputing*, 214, 546–554.

Hung, G. M., John, N. W., Hancock, C. and Hoshi, T., 2014. Using and validating airborne ultrasound as a tactile interface within medical training simulators. *International Symposium on Biomedical Simulation*, Springer, 30–39.

Hung, J.-L. and Zhang, K., 2012. Examining mobile learning trends 2003–2008: A categorical meta-trend analysis using text mining techniques. *Journal of Computing in Higher education*, 24 (1), 1–17.

Ieiri, S., Uemura, M., Konishi, K., Souzaki, R., Nagao, Y., Tsutsumi, N., Akahoshi, T., Ohuchida, K., Ohdaira, T., Tomikawa, M. et al., 2012. Augmented reality navigation system for laparoscopic splenectomy in children based on preoperative ct image using optical tracking device. *Pediatric surgery international*, 28 (4), 341–346.

Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A. A., 2017. Image-to-image translation with conditional adversarial networks. *CVPR*.

Jaderberg, M., Simonyan, K., Zisserman, A. and kavukcuoglu, k., 2015. Spatial transformer networks. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, eds., *Advances in Neural Information Processing Systems 28*, Curran Associates, Inc., 2017–2025. URL `http://papers.nips.cc/paper/5854-spatial-transformer-networks.pdf`.

Jeon, S. and Choi, S., 2010. Stiffness modulation for haptic augmented reality: Extension to 3d interaction. *Haptics Symposium, 2010 IEEE*, IEEE, 273–280.

Jeon, S., Choi, S. and Harders, M., 2012. Rendering virtual tumors in real tissue mock-ups using haptic augmented reality. *Haptics, IEEE Transactions on*, 5 (1), 77–84.

Jeon, S., Knoerlein, B., Harders, M. and Choi, S., 2010a. Haptic simulation of breast cancer palpation: A case study of haptic augmented reality. *Mixed and Augmented Reality (ISMAR), 2010 9th IEEE International Symposium on*, IEEE, 237–238.

Jeon, S., Knoerlein, B., Harders, M., Han, G. and Choi, S., 2010b. Breast cancer palpation system using haptic augmented reality. *Mixed and Augmented Reality (ISMAR), 2010 9th IEEE International Symposium on*, IEEE, 308–308.

Jeroudi, O. M., Christakopoulos, G., Christopoulos, G., Kotsia, A., Kypreos, M. A., Rangan, B. V., Banerjee, S. and Brilakis, E. S., 2015. Accuracy of remote electrocardiogram interpretation with the use of google glass technology. *The American journal of cardiology*, 115 (3), 374–377.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S. and Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. *Proceedings of the 22Nd ACM International Conference on Multimedia*, New York, NY, USA: ACM, MM '14, 675–678. URL `http://doi.acm.org/10.1145/2647868.2654889`.

Jiang, J., Xing, Y., Wang, S. and Liang, K., 2017. Evaluation of robotic surgery skills using dynamic time warping. *Computer Methods and Programs in Biomedicine*, 152 (Supplement C), 71 – 83. URL `http://www.sciencedirect.com/science/article/pii/S0169260716308513`.

Johnson, A., 1999. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 21 (5), 433–449.

Johnson, A. S. and Sun, Y., 2013. Exploration of spatial augmented reality on person. *Virtual Reality (VR), 2013 IEEE*, IEEE, 59–60.

Juanes, J. A., Hernández, D., Ruisoto, P., García, E., Villarrubia, G. and Prats, A., 2014. Augmented reality techniques, using mobile devices, for learning human anatomy. *Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality*, ACM, 7–11.

Kang, X., Azizian, M., Wilson, E., Wu, K., Martin, A. D., Kane, T. D., Peters, C. A., Cleary, K. and Shekhar, R., 2014. Stereoscopic augmented reality for laparoscopic surgery. *Surgical endoscopy*, 28 (7), 2227–2235.

Kato, H. and Billinghurst, M., 1999. Marker tracking and hmd calibration for a video-based augmented reality conferencing system. *Augmented Reality,*

1999.(IWAR'99) Proceedings. 2nd IEEE and ACM International Workshop on, IEEE, 85–94.

Kazhdan, M. and Hoppe, H., 2013. Screened poisson surface reconstruction. *ACM Trans. Graph.*, 32 (3), 29:1–29:13. URL `http://doi.acm.org/10.1145/2487228.2487237`.

Ke, Y. and Sukthankar, R., 2004. Pca-sift: A more distinctive representation for local image descriptors. *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, IEEE, volume 2, II–506.

Kendall, A., Martirosyan, H., Dasgupta, S. and Henry, P., 2017. End-to-end learning of geometry and context for deep stereo regression. *2017 IEEE International Conference on Computer Vision (ICCV)*, 66–75.

Khan, M. F., Dogan, S., Maataoui, A., Wesarg, S., Gurung, J., Ackermann, H., Schiemann, M., Wimmer-Greinecker, G. and Vogl, T. J., 2006. Navigation-based needle puncture of a cadaver using a hybrid tracking navigational system. *Investigative radiology*, 41 (10), 713–720.

Kilgus, T., Heim, E., Haase, S., Prüfer, S., Müller, M., Seitel, A., Fangerau, M., Wiebe, T., Iszatt, J., Schlemmer, H.-P. et al., 2015. Mobile markerless augmented reality and its application in forensic medicine. *International journal of computer assisted radiology and surgery*, 10 (5), 573–586.

Kim, H., Takahashi, I., Yamamoto, H., Kai, T., Maekawa, S. and Naemura, T., 2013. *MARIO: Mid-Air Augmented RealityInteraction with Objects*, Cham: Springer International Publishing. 560–563.

Kim, J.-H., Bartoli, A., Collins, T. and Hartley, R., 2012. Tracking by detection for interactive image augmentation in laparoscopy. *Lecture Notes in Computer Science*, 246?255.

Kim, S.-C., Han, B.-K., Seo, J. and Kwon, D.-S., 2014. Haptic interaction with virtual geometry on robotic touch surface. *SIGGRAPH Asia 2014 Emerging Technologies*, ACM, 8.

Kindermann, R. and Snell, J. L., 1980. *Markov random fields and their applications*, volume 1. American Mathematical Society.

Klein, G. and Murray, D., 2007. Parallel tracking and mapping for small ar workspaces. *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*, IEEE, 225–234.

Knook, M., Rosmalen, A., Yoder, B., Kleinrensink, G., Snijders, C., Looman, C. and Steensel, C., 2001. Optimal mesh size for endoscopic inguinal herniarepair. *Surgical Endoscopy*, 15 (12), 1471–1477. URL `https://doi.org/10.1007/s00464-001-0048-9`.

Kotwal, A., Bhalodia, R. and Awate, S. P., 2016. Joint desmoking and denoising of laparoscopy images. *IEEE ISBI*.

Krähenbühl, P. and Koltun, V., 2011. Efficient inference in fully connected crfs with gaussian edge potentials. J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems 24*, Curran Associates, Inc., 109–117. URL `http://papers.nips.cc/paper/4296-efficient-inference-in-fully-connected-crfs-with-gaussian-edge-potentia pdf`.

Kramers, M., Armstrong, R., Bakhshmand, S. M., Fenster, A., de Ribaupierre, S. and Eagleson, R., 2014. Evaluation of a mobile augmented reality application for image guidance of neurosurgical interventions. *Stud Health Technol Inform*, 196, 204–8.

Kratzer, W., Fritz, V., Mason, R. A., Haenle, M. M., Kaechele, V. and , R. S. G., 2003. Factors affecting liver size: a sonographic survey of 2080 subjects. *J Ultrasound Med*, 22 (11), 1155–1161.

Krizhevsky, A., Sutskever, I. and Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems*, USA: Curran Associates Inc., NIPS'12, 1097–1105. URL http://dl.acm.org/citation.cfm?id=2999134.2999257.

Kuhlemann, I., Kleemann, M., Jauer, P., Schweikard, A. and Ernst, F., 2017. Towards x-ray free endovascular interventions – using hololens for on-line holographic visualisation. *Healthcare Technology Letters*, 4 (5), 184–187.

Kumar, A., Wang, Y.-Y., Wu, C.-J., Liu, K.-C. and Wu, H.-S., 2014. Stereoscopic visualization of laparoscope image using depth information from 3D model. *Computer Methods and Programs in Biomedicine*, 113 (3), 862–868. URL http://dx.doi.org/10.1016/j.cmpb.2013.12.013.

Kuznietsov, Y., St'ckler, J. and Leibe, B., 2017. Semi-supervised deep learning for monocular depth map prediction. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2215–2223.

Kwartowitz, D. M., Rettmann, M. E., Holmes III, D. R. and Robb, R. A., 2010. A novel technique for analysis of accuracy of magnetic tracking systems used in image guided surgery. *SPIE Medical Imaging*, International Society for Optics and Photonics, 76251L–76251L.

Ladick', L., Russell, C., Kohli, P. and Torr, P. H. S., 2009. Associative hierarchical crfs for object class image segmentation. *2009 IEEE 12th International Conference on Computer Vision*, 739–746.

Ladický, L., Shi, J. and Pollefeys, M., 2014. Pulling things out of perspective. *2014 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, Washington, DC, USA: IEEE Computer Society, CVPR '14, 89–96. URL http://dx.doi.org/10.1109/CVPR.2014.19.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F. and Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks. *3D Vision (3DV), 2016 Fourth International Conference on*, 239–248.

Lamata, P., Freudenthal, A., Cano, A., Kalkofen, D., Schmalstieg, D., Naerum, E., Samset, E., Gómez, E. J., Sánchez-Margallo, F. M., Furtado, H. et al., 2010. *Augmented reality for minimally invasive surgery: overview and some recent advances*. INTECH Open Access Publisher.

Lambert, L., Ahmed, S. Z., Hachicha, K., Pinna, A. and Garda, P., 2016. High frame rate medical quality video compression for tele-EEG. *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. URL `http://dx.doi.org/10.1109/EMBC.2016.7591946`.

Lee, J.-D., Huang, C.-H., Huang, T.-C., Hsieh, H.-Y. and Lee, S.-T., 2012. Medical augment reality using a markerless registration framework. *Expert Systems with Applications*, 39 (5), 5286–5294.

Lee, S., Lee, J., Lee, A., Park, N., Song, S., Seo, A., Lee, H., Kim, J.-I. and Eom, K., 2013. Augmented reality intravenous injection simulator based 3d medical imaging for veterinary medicine. *The Veterinary Journal*, 196 (2), 197–202.

Lee, T. and Hollerer, T., 2008. Hybrid feature tracking and user interaction for markerless augmented reality. *Virtual Reality Conference, 2008. VR'08. IEEE*, IEEE, 145–152.

Leutenegger, S., Chli, M. and Siegwart, R. Y., 2011. Brisk: Binary robust invariant scalable keypoints. *Computer Vision (ICCV), 2011 IEEE International Conference on*, IEEE, 2548–2555.

Levin, D., 2004. Mesh-independent surface interpolation. *Mathematics and Visualization*, 37–49.

Li, B., Peng, X., Wang, Z., Xu, J. and Feng, D., 2017a. AOD-net: All-in-one dehazing network. *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE.

Li, B., Shen, C., Dai, Y., van den Hengel, A. and He, M., 2015. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1119–1127.

Li, J., Klein, R. and Yao, A., 2017b. A two-streamed network for estimating fine-scaled depth maps from single rgb images. *2017 IEEE International Conference on Computer Vision (ICCV)*, 3392–3400.

Li, L.-L., Ding, G., Feng, N., Wang, M.-H. and Ho, Y.-S., 2009. Global stem cell research trend: Bibliometric analysis as a tool for mapping of trends from 1991 to 2006. *Scientometrics*, 80 (1), 39–58.

Liao, H., Inomata, T., Sakuma, I. and Dohi, T., 2010. 3-D augmented reality for MRI-guided surgery using integral videography autostereoscopic image overlay. *Biomedical Engineering, IEEE Transactions on*, 57 (6), 1476–1486.

Liao, H., Iwahara, M., Hata, N. and Dohi, T., 2004. High-quality integral videography using a multiprojector. *Optics Express*, 12 (6), 1067–1076.

Lin, B., Johnson, A., Qian, X., Sanchez, J. and Sun, Y., 2013. Simultaneous tracking, 3D reconstruction and deforming point detection for stereoscope guided surgery. *Lecture Notes in Computer Science*, 35–44.

Lin, J., Clancy, N. T., Hu, Y., Qi, J., Tatla, T., Stoyanov, D., Maier-Hein, L. and Elson, D. S., 2017. Endoscopic depth measurement and super-spectral-resolution imaging. *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017*.

Lin, J., Nishino, H., Kagawa, T. and Utsumiya, K., 2012. A method of two-handed gesture interactions with applications based on commodity devices. *Computers &*

Mathematics with Applications, 63 (2), 448?457. URL http://dx.doi.org/10.1016/j.camwa.2011.07.052.

Ling, H. and Jacobs, D., 2005. Deformation invariant image matching. *Tenth IEEE International Conference on Computer Vision (ICCV 05) Volume 1*, Institute of Electrical & Electronics Engineers (IEEE). URL http://dx.doi.org/10.1109/ICCV.2005.67.

Liu, F., Shen, C., Lin, G. and Reid, I., 2016. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38 (10), 2024–2039.

Liu, M., Salzmann, M. and He, X., 2014. Discrete-continuous depth estimation from a single image. *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 716–723.

London, I. C., 2016. Hamlyn centre laparoscopic / endoscopic video datasets. URL http://hamlyn.doc.ic.ac.uk/vision/, [Accessed 6 Nov. 2016].

Lotte, F., Faller, J., Guger, C., Renard, Y., Pfurtscheller, G., Lécuyer, A. and Leeb, R., 2012. Combining bci with virtual reality: towards new applications and improved bci. *Towards Practical Brain-Computer Interfaces*, Springer, 197–220.

Loukas, C., Lahanas, V. and Georgiou, E., 2013. An integrated approach to endoscopic instrument tracking for augmented reality applications in surgical simulation training. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 9 (4), e34–e51.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60 (2), 91–110.

Luo, X., McLeod, A. J., Pautler, S. E., Schlachta, C. M. and Peters, T. M., 2017. Vision-based surgical field defogging. *IEEE Transactions on Medical Imaging*, 36 (10), 2021–2030.

Luo, Y., Ren, J., Lin, M., Pang, J., Sun, W., Li, H. and Lin, L., 2018. Single view stereo matching. *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ma, M., Fallavollita, P., Seelbach, I., Heide, A. M., Euler, E., Waschke, J. and Navab, N., 2015. Personalized augmented reality for anatomy education. *Clinical Anatomy*.

Mahmoud, N., Cirauqui, I., Hostettler, A., Doignon, C., Soler, L., Marescaux, J. and Montiel, J., 2016. Orbslam-based endoscope tracking and 3d reconstruction. *MICCAI 2016 CARE*.

Maletsky, L. P., Sun, J. and Morton, N. A., 2007. Accuracy of an optical active-marker system to track the relative motion of rigid bodies. *Journal of biomechanics*, 40 (3), 682–685.

Marescaux, J. and Diana, M., 2015. Next step in minimally invasive surgery: hybrid image-guided surgery. *Journal of pediatric surgery*, 50 (1), 30–36.

Markelj, P., Tomaževič, D., Likar, B. and Pernuš, F., 2012. A review of 3d/2d registration methods for image-guided interventions. *Medical image analysis*, 16 (3), 642–661.

Marton, Z. C., Rusu, R. B. and Beetz, M., 2009. On fast surface reconstruction methods for large and noisy point clouds. *2009 IEEE International Conference on Robotics and Automation*, 3218–3223.

Maurer Jr, C. R., Fitzpatrick, J. M., Wang, M. Y., Galloway, R. L., Maciunas, R. J. and Allen, G. S., 1997. Registration of head volume images using implantable fiducial markers. *Medical Imaging, IEEE Transactions on*, 16 (4), 447–462.

Mayer, N., Ilg, E., H'usser, P., Fischer, P., Cremers, D., Dosovitskiy, A. and Brox, T., 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4040–4048.

McCartney, E. J. and Hall, F. F., 1977. Optics of the atmosphere: Scattering by molecules and particles. *Physics Today*, 30 (5), 76–77.

McCormac, J., Handa, A., Davison, A. and Leutenegger, S., 2017. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 4628–4635.

Meng, G., Wang, Y., Duan, J., Xiang, S. and Pan, C., 2013a. Efficient image dehazing with boundary constraint and contextual regularization. *2013 IEEE International Conference on Computer Vision*, IEEE.

Meng, G. C., Shahzad, A., Saad, N., Malik, A. S. and Meriaudeau, F., 2015. Prototype design for wearable veins localization system using near infrared imaging technique. *Signal Processing & Its Applications (CSPA), 2015 IEEE 11th International Colloquium on*, IEEE, 112–115.

Meng, M., Fallavollita, P., Blum, T., Eck, U., Sandor, C., Weidert, S., Waschke, J. and Navab, N., 2013b. Kinect for interactive ar anatomy learning. *Mixed and Augmented Reality (ISMAR), 2013 IEEE International Symposium on*, IEEE, 277–278.

Mercier-Ganady, J., Lotte, F., Loup-Escande, E., Marchal, M. and Lécuyer, A., 2014. The mind-mirror: See your brain in action in your head using eeg and augmented reality. *Virtual Reality (VR), 2014 iEEE*, IEEE, 33–38.

Michael, K., Bolitho, M. and Hoppe, H., 2006. Poisson surface reconstruction. *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006.

Microsoft, 2017. Ball pit - microsoft store. URL `https://www.microsoft.com/en-us/store/p/ball-pit/9nblggh4wssp`.

Microsoft, 2018. Microsoft hololens. https://www.microsoft.com/microsoft-hololens. [Accessed 29 4 2016].

Mikolajczyk, K. and Schmid, C., 2005. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27 (10), 1615–1630.

Milgram, P. and Kishino, F., 1994. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems*, 77 (12), 1321–1329.

Mirota, D. J., Ishii, M. and Hager, G. D., 2011. Vision-based navigation in image-guided interventions. *Annual review of biomedical engineering*, 13, 297–319.

Mistry, M., Roach, V. A. and Wilson, T. D., 2013. Application of stereoscopic visualization on surgical skill acquisition in novices. *Journal of Surgical Education*, 70 (5), 563 – 570. URL `http://www.sciencedirect.com/science/article/pii/S1931720413001165`.

Mountney, P., Fallert, J., Nicolau, S., Soler, L. and Mewes, P. W., 2014. An augmented reality framework for soft tissue surgery. *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2014*.

Mountney, P., Lo, B., Thiemjarus, S., Stoyanov, D. and Zhong-Yang, G., 2007. A probabilistic framework for tracking deformable soft tissue in minimally invasive surgery. *Med Image Comput Comput Assist Interv*, 10 (Pt 2), 34–41.

Mountney, P., Stoyanov, D., Davison, A. and Yang, G.-Z., 2006. Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery. *Med Image Comput Comput Assist Interv*, 9 (Pt 1), 347–354.

Mountney, P. and Yang, G.-Z., 2008. Soft tissue tracking for minimally invasive surgery: learning local deformation online. *Med Image Comput Comput Assist Interv*, 11 (Pt 2), 364–372.

Mountney, P. and Yang, G.-Z., 2009. Dynamic view expansion for minimally invasive surgery using simultaneous localization and mapping. *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. URL `http://dx.doi.org/10.1109/IEMBS.2009.5333939`.

Mountney, P. and Yang, G.-Z., 2010a. Motion compensated slam for image guided surgery. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, Springer, 496–504.

Mountney, P. and Yang, G.-Z., 2010b. Motion compensated slam for image guided surgery. *Medical Image Computing and Computer-Assisted Intervention MICCAI 2010*.

Mousavian, A., Pirsiavash, H. and Košecká, J., 2016. Joint semantic segmentation and depth estimation with deep convolutional networks. *3D Vision (3DV), 2016 Fourth International Conference on*, IEEE, 611–619.

Muensterer, O. J., Lacher, M., Zoeller, C., Bronstein, M. and Kübler, J., 2014. Google glass in pediatric surgery: an exploratory study. *International Journal of Surgery*, 12 (4), 281–289.

Mur-Artal, R., Montiel, J. M. M. and Tardós, J. D., 2015. Orb-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31 (5), 114–1163.

Murali, A., Sen, S., Kehoe, B., Garg, A., McFarland, S., Patil, S., Boyd, W. D., Lim, S., Abbeel, P. and Goldberg, K., 2015. Learning by observation for surgical subtasks: Multilateral cutting of 3d viscoelastic and 2d orthotropic tissue phantoms. *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 1202–1209.

Naimark, L. and Foxlin, E., 2002. Circular data matrix fiducial system and robust image processing for a wearable vision-inertial self-tracker. *Proceedings of the 1st International Symposium on Mixed and Augmented Reality*, IEEE Computer Society, 27.

Nakata, N., Suzuki, N., Hattori, A., Hirai, N., Miyamoto, Y. and Fukuda, K., 2012. Informatics in radiology: intuitive user interface for 3d image manipulation using

augmented reality and a smartphone as a remote control. *Radiographics*, 32 (4), E169–E174.

Narasimhan, S. and Nayar, S., 2003. Contrast restoration of weather degraded images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25 (6), 713–724.

Nayar, S. and Narasimhan, S., 1999. Vision in bad weather. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, IEEE.

Neira, J. and Tardós, J. D., 2001. Data association in stochastic mapping using the joint compatibility test. *Robotics and Automation, IEEE Transactions on*, 17 (6), 890–897.

Newcombe, R. A., Fox, D. and Seitz, S. M., 2015. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. *Proceedings of the IEEE CVPR*, 343–352.

Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohi, P., Shotton, J., Hodges, S. and Fitzgibbon, A., 2011a. Kinectfusion: Real-time dense surface mapping and tracking. *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, IEEE, 127–136. URL `http://dx.doi.org/10.1109/CVPR.2015.7298631`.

Newcombe, R. A., Lovegrove, S. J. and Davison, A. J., 2011b. Dtam: Dense tracking and mapping in real-time. *2011 International Conference on Computer Vision*, 2320–2327.

Nicolau, S., Garcia, A., Pennec, X., Soler, L. and Ayache, N., 2005. An augmented reality system to guide radio-frequency tumour ablation. *Computer animation and virtual worlds*, 16 (1), 1–10.

Nicolau, S., Soler, L., Mutter, D. and Marescaux, J., 2011. Augmented reality in laparoscopic surgical oncology. *Surgical oncology*, 20 (3), 189–201.

Nishino, K., Kratz, L. and Lombardi, S., 2011. Bayesian defogging. *International Journal of Computer Vision*, 98 (3), 263–278.

Noll, C., Häussermann, B., von Jan, U., Raap, U. and Albrecht, U.-V., 2014. Demo: Mobile augmented reality in medical education: An application for dermatology. *Proceedings of the 2014 Workshop on Mobile Augmented Reality and Robotic Technology-based Systems*, New York, NY, USA: ACM, MARS '14, 17–18. URL `http://doi.acm.org/10.1145/2609829.2609833`.

North, C., Endert, A. and Fiaux, P., 2012. Semantic interaction for sensemaking: Inferring analytical reasoning for model steering. *IEEE Transactions on Visualization & Computer Graphics*, 18, 2879–2888. URL `doi.ieeecomputersociety.org/10.1109/TVCG.2012.260`.

Nuernberger, B., Ofek, E., Benko, H. and Wilson, A. D., 2016. Snaptoreality: Aligning augmented reality to the real world. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, CHI '16, 1233–1244. URL `http://doi.acm.org/10.1145/2858036.2858250`.

Nützi, G., Weiss, S., Scaramuzza, D. and Siegwart, R., 2011. Fusion of imu and vision for absolute scale estimation in monocular slam. *Journal of intelligent & robotic systems*, 61 (1), 287–299.

Oktay, O., Zhang, L., Mansi, T., Mountney, P., Mewes, P., Nicolau, S., Soler, L. and Chefdhótel, C., 2013. Biomechanically driven registration of pre-to intra-operative 3d images for laparoscopic surgery. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, Springer, 1–9.

Olsson, T. and Salo, M., 2011. Online user survey on current mobile augmented reality applications. *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, IEEE, 75–84.

Plantefève, R., Peterlik, I., Haouchine, N. and Cotin, S., 2016. Patient-specific

biomechanical modeling for guidance during minimally-invasive hepatic surgery. *Annals of biomedical engineering*, 44 (1), 139–153.

Planteféve, R., Peterlik, I., Haouchine, N. and Cotin, S., 2016. Patient-specific biomechanical modeling for guidance during minimally-invasive hepatic surgery. *Ann Biomed Eng*, 44 (1), 139–153. URL `http://dx.doi.org/10.1007/s10439-015-1419-z`.

Prados, E. and Faugeras, O., 2005. Shape from shading: A well-posed problem? *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR?05)*. URL `http://dx.doi.org/10.1109/CVPR.2005.319`.

PTC, 2018. Vuforia. http://www.vuforia.com/. [Accessed 29 4 2016].

Puerto, G. A. and Mariottini, G.-L., 2012. A comparative study of correspondence-search algorithms in mis images. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, Springer, 625–633.

Puerto-Souza, G. A. and Mariottini, G.-L., 2013. A fast and accurate feature-matching algorithm for minimally-invasive endoscopic images. *Medical Imaging, IEEE Transactions on*, 32 (7), 1201–1214.

Raab, F. H., Blood, E. B., Steiner, T. O. and Jones, H. R., 1979. Magnetic position and orientation tracking system. *Aerospace and Electronic Systems, IEEE Transactions on*, (5), 709–718.

Rantakari, J., Colley, A. and Häkkilä, J., 2015. Exploring ar poster as an interface to personal health data. *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*, ACM, 422–425.

Raskar, R., Welch, G. and Fuchs, H., 1998. Spatially augmented reality. *First IEEE Workshop on Augmented Reality (IWAR'98)*, Citeseer, 11–20.

Rautek, P., Bruckner, S. and Groller, E., 2007. Semantic layers for illustrative volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 13 (6), 1336–1343.

Rekimoto, J. and Ayatsuka, Y., 2000. Cybercode: designing augmented reality environments with visual tags. *Proceedings of DARE 2000 on Designing augmented reality environments*, ACM, 1–10.

Reuters, T., 2016. Web of knowledge - real facts - ip & science - thomson reuters. http://wokinfo.com/citationconnection/realfacts/.

Rolland, J. P. and Fuchs, H., 2000. Optical versus video see-through head-mounted displays in medical visualization. *Presence*, 9 (3), 287–309.

Ronneberger, O., Fischer, P. and Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*.

Rublee, E., Rabaud, V., Konolige, K. and Bradski, G., 2011. Orb: an efficient alternative to sift or surf. *Computer Vision (ICCV), 2011 IEEE International Conference on*, IEEE, 2564–2571.

Rünz, M. and Agapito, L., 2017. Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 4471–4478.

Saga, S. and Deguchi, K., 2012. Lateral-force-based 2.5-dimensional tactile display for touch screen. *Haptics Symposium (HAPTICS), 2012 IEEE*, IEEE, 15–22.

Salas-Moreno, R. F., Glocken, B., Kelly, P. H. J. and Davison, A. J., 2014. Dense planar slam. *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 157–164.

Samosky, J. T., Nelson, D. A., Wang, B., Bregman, R., Hosmer, A., Mikulis, B. and Weaver, R., 2012. Bodyexplorerar: enhancing a mannequin medical simulator with sensing and projective augmented reality for exploring dynamic anatomy and physiology. *Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction*, ACM, 263–270.

Sarayeddine, K. and Mirza, K., 2013. Key challenges to affordable see-through wearable displays: the missing link for mobile ar mass deployment. *SPIE Defense, Security, and Sensing*, International Society for Optics and Photonics, 87200D–87200D.

Saunders, J. A. and Backus, B. T., 2006. The accuracy and reliability of perceived depth from linear perspective as a function of image size. *Journal of Vision*, 6 (9), 7–7.

Saxena, A., Chung, S. H. and Ng, A. Y., 2006. Learning depth from single monocular images. Y. Weiss, B. Schölkopf and J. C. Platt, eds., *Advances in Neural Information Processing Systems 18*, MIT Press, 1161–1168. URL `http://papers.nips.cc/paper/2921-learning-depth-from-single-monocular-images.pdf`.

Saxena, A., Chung, S. H. and Ng, A. Y., 2008. 3-d depth reconstruction from a single still image. *International Journal of Computer Vision*, 76 (1), 53–69. URL `https://doi.org/10.1007/s11263-007-0071-y`.

Saxena, A., Schulte, J. and Ng, A. Y., 2007. Depth estimation using monocular and stereo cues. *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., IJCAI'07, 2197–2203. URL `http://dl.acm.org/citation.cfm?id=1625275.1625630`.

Saxena, A., Sun, M. and Ng, A. Y., 2009. Make3d: Learning 3D scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31 (5), 824–840.

Scharstein, D., Szeliski, R. and Zabih, R., 2001. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, 131–140.

Shelhamer, E., Long, J. and Darrell, T., 2017. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (4), 640–651.

Shi, C., Becker, B. C. and Riviere, C. N., 2012. Inexpensive monocular pico-projector-based augmented reality display for surgical microscope. *Computer-Based Medical Systems (CBMS), 2012 25th International Symposium on*, IEEE, 1–6.

Shotton, J., Winn, J., Rother, C. and Criminisi, A., 2009. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81 (1), 2–23.

Simonyan, K. and Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*. URL `https://arxiv.org/abs/1409.1556`.

Smith, R. C. and Cheeseman, P., 1986. On the representation and estimation of spatial uncertainty. *The international journal of Robotics Research*, 5 (4), 56–68.

Soeiro, J., Cláudio, A. P., Carmo, M. B. and Ferreira, H. A., 2015. Visualizing the brain on a mixed reality smartphone application. *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, IEEE, 5090–5093.

Song, T., Yang, C., Dianat, O. and Azimi, E., 2018. Endodontic guided treatment using augmented reality on a head-mounted display system. *Healthcare Technology Letters*, 5 (5), 201–207.

Sousa Silva, E. and Formico Rodrigues, M. A., 2015. Gesture interaction and evaluation using the leap motion for medical visualization. *Virtual and Augmented Reality (SVR), 2015 XVII Symposium on*, IEEE, 160–169.

Spillmann, J., Tuchschmid, S. and Harders, M., 2013. Adaptive space warping to enhance passive haptics in an arthroscopy surgical simulator. *Visualization and Computer Graphics, IEEE Transactions on*, 19 (4), 626–633.

Stefan, P., Wucherer, P., Oyamada, Y., Ma, M., Schoch, A., Kanegae, M., Shimizu, N., Kodera, T., Cahier, S., Weigl, M. et al., 2014. An ar edutainment system supporting bone anatomy learning. *Virtual Reality (VR), 2014 iEEE*, IEEE, 113–114.

Stoyanov, D., 2012. Stereoscopic scene flow for robotic assisted minimally invasive surgery. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, Springer, 479–486.

Stoyanov, D., Darzi, A. and Yang, G. Z., 2004. Dense 3d depth recovery for soft tissue deformation during robotically assisted laparoscopic surgery. *Medical Image Computing and Computer-Assisted Intervention MICCAI 2004*.

Stoyanov, D., Darzi, A. and Yang, G. Z., 2005a. A practical approach towards accurate dense 3d depth recovery for robotic laparoscopic surgery. *Comput Aided Surg*, 10 (4), 199–208. URL `http://dx.doi.org/10.3109/10929080500230379`.

Stoyanov, D., Mylonas, G. P., Deligianni, F., Darzi, A. and Yang, G. Z., 2005b. Soft-tissue motion tracking and structure estimation for robotic assisted mis procedures. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*, Springer, 139–146.

Stoyanov, D., Scarzanella, M. V., Pratt, P. and Yang, G.-Z., 2010. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. *Medical Image Computing and Computer-Assisted Intervention MICCAI 2010*.

Stredney, D. and Weghorst, S., 1998. The virtual retinal display: a new technology for virtual reality and augmented vision in medicine. *Medicine Meets Virtual Reality: Art, Science, Technology: Healthcare (R) Evolution*, 50, 252.

Su, L.-M., Vagvolgyi, B. P., Agarwal, R., Reiley, C. E., Taylor, R. H. and Hager, G. D., 2009. Augmented reality during robot-assisted laparoscopic partial nephrectomy: Toward real-time 3D-CT to stereoscopic video registration. *Urology*, 73 (4), 896–900.

Sulami, M., Glatzer, I., Fattal, R. and Werman, M., 2014. Automatic recovery of the atmospheric light in hazy images. *2014 IEEE International Conference on Computational Photography (ICCP)*, IEEE.

Sutherland, C., Hashtrudi-Zaad, K., Sellens, R., Abolmaesumi, P. and Mousavi, P., 2013. An augmented reality haptic training simulator for spinal needle procedures. *Biomedical Engineering, IEEE Transactions on*, 60 (11), 3009–3018.

Sutherland, I. E., 1968. A head-mounted three dimensional display. *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, ACM, 757–764.

Tan, R. T., 2008. Visibility in bad weather from a single image. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE.

Tanaka, T., 2010. Clinical application for assistive engineering–mixed reality rehabilitation. *Journal of Medical and Biological Engineering*, 31 (4), 277–282.

Tang, K., Yang, J. and Wang, J., 2014. Investigating haze-relevant features in a learning framework for image dehazing. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE.

Tateno, K., Tombari, F., Laina, I. and Navab, N., 2017. Cnn-slam: Real-time dense monocular slam with learned depth prediction. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6565–6574.

Tchaka, K., Pawar, V. M. and Stoyanov, D., 2017. Chromaticity based smoke removal in endoscopic images. M. A. Styner and E. D. Angelini, eds., *Medical Imaging 2017: Image Processing*, SPIE.

Tomasi, C. and Kanade, T., 1991. *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh.

Totz, J., Mountney, P., Stoyanov, D. and Yang, G.-Z., 2011. Dense surface reconstruction for enhanced navigation in mis. *Med Image Comput Comput Assist Interv*, 14 (Pt 1), 89–96.

Tsui, C., Klein, R. and Garabrant, M., 2013. Minimally invasive surgery: national trends in adoption and future directions for hospital strategy. *Surgical endoscopy*, 27, 2253–2257.

Tully, J., Dameff, C., Kaib, S. and Moffitt, M., 2015. Recording medical students' encounters with standardized patients using google glass: Providing end-of-life clinical education. *Academic Medicine*, 90 (3), 314–316.

Turan, M., Almalioglu, Y., Araujo, H., Konukoglu, E. and Sitti, M., 2017. A Non-Rigid Map Fusion-Based RGB-Depth SLAM Method for Endoscopic Capsule Robots. *ArXiv e-prints*.

Turini, G., Condino, S., Parchi, P. D., Viglialoro, R. M., Piolanti, N., Gesi, M., Ferrari, M. and Ferrari, V., 2018. A microsoft hololens mixed reality surgical simulator for patient-specific hip arthroplasty training. L. T. De Paolis and P. Bourdot, eds., *Augmented Reality, Virtual Reality, and Computer Graphics*, Cham: Springer International Publishing, 201–210.

Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M. and Padoy, N., 2017. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36 (1), 86–97.

Ulmer, B. C., 2008. The hazards of surgical smoke. *AORN Journal*, 87 (4), 721–738.

Van Krevelen, D. and Poelman, R., 2010. A survey of augmented reality technologies, applications and limitations. *International Journal of Virtual Reality*, 9 (2), 1.

Vanga, S. K., Singh, A., Vagadia, B. H. and Raghavan, V., 2015. Global food allergy research trend: a bibliometric analysis. *Scientometrics*, 105 (1), 203–213.

Velayutham, V., Fuks, D., Nomi, T., Kawaguchi, Y. and Gayet, B., 2016. 3d visualization reduces operating time when compared to high-definition 2d in laparoscopic liver resection: a case-matched study. *Surgical endoscopy*, 30 (1), 147–153.

Vigh, B., Müller, S., Ristow, O., Deppe, H., Holdstock, S., den Hollander, J., Navab, N., Steiner, T. and Hohlweg-Majert, B., 2014. The use of a head-mounted display in oral implantology: a feasibility study. *International journal of computer assisted radiology and surgery*, 9 (1), 71–78.

Virag, I., Stoicu-Tivadar, L. and Amaricai, E., 2014. Browser-based medical visualization system. *Applied Computational Intelligence and Informatics (SACI), 2014 IEEE 9th International Symposium on*, IEEE, 355–359.

Visentini-Scarzanella, M., Stoyanov, D. and Yang, G.-Z., 2012. Metric depth recovery from monocular images using shape-from-shading and specularities. *2012 19th IEEE International Conference on Image Processing*. URL `http://dx.doi.org/10.1109/ICIP.2012.6466786`.

Wagner, D. and Schmalstieg, D., 2006. Handheld augmented reality displays. *Virtual Reality Conference, 2006*, IEEE, 321–321.

Wagner, O. J., Hagen, M., Kurmann, A., Horgan, S., Candinas, D. and Vorburger, S. A., 2012. Three-dimensional vision enhances task performance independently of the surgical method. *Surg Endosc*, 26 (10), 2961–2968. URL `http://dx.doi.org/10.1007/s00464-012-2295-3`.

Wakabayashi, D., 2015. Apple buys german augmented-reality firm metaio. http://blogs.wsj.com/digits/2015/05/28/apple-buys-german-augmented-reality-firm-metaio/. [Accessed 29 4 2016].

Wang, C., Alaya Cheikh, F., Kaaniche, M. and Elle, O. J., 2018. A Smoke Removal Method for Laparoscopic Images. *ArXiv e-prints*.

Wang, H., Wang, F., Leong, A. P. Y., Xu, L., Chen, X. and Wang, Q., 2015a. Precision insertion of percutaneous sacroiliac screws using a novel augmented reality-based navigation system: a pilot study. *International orthopaedics*, 1–7.

Wang, J., Qi, L. and Meng, M. Q.-H., 2015b. Robot-assisted occlusion avoidance for

surgical instrument optical tracking system. *Information and Automation, 2015 IEEE International Conference on*, IEEE, 375–380.

Wang, J., Suenaga, H., Hoshi, K., Yang, L., Kobayashi, E., Sakuma, I. and Liao, H., 2014. Augmented reality navigation with automatic marker-free image registration using 3-d image overlay for dental surgery. *Biomedical Engineering, IEEE Transactions on*, 61 (4), 1295–1304.

Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B. and Yuille, A., 2015c. Towards unified depth and semantic prediction from a single image. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2800–2809.

Weese, J., Penney, G. P., Desmedt, P., Buzug, T. M., Hill, D. L. and Hawkes, D. J., 1997. Voxel-based 2-d/3-d registration of fluoroscopy images and ct scans for image-guided surgery. *Information Technology in Biomedicine, IEEE Transactions on*, 1 (4), 284–293.

Weld, K. J., Dryer, S., Ames, C. D., Cho, K., Hogan, C., Lee, M., Biswas, P. and Landman, J., 2007. Analysis of surgical smoke produced by various energy-based instruments and effect on laparoscopic visibility. *Journal of endourology*, 21, 347–351.

Whelan, T., Johannsson, H., Kaess, M., Leonard, J. J. and McDonald, J., 2013. Robust real-time visual odometry for dense rgb-d mapping. *2013 IEEE International Conference on Robotics and Automation*, 5724–5731.

Widmer, A., Schaer, R., Markonis, D. and Muller, H., 2014. Facilitating medical information search using google glass connected to a content-based medical image retrieval system. *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, IEEE, 4507–4510.

Wieczorek, M., Aichert, A., Kutter, O., Bichlmeier, C., Landes, J., Heining, S. M., Euler, E. and Navab, N., 2010. Gpu-accelerated rendering for medical augmented

reality in minimally-invasive procedures. *Bildverarbeitung für die Medizin*, 102–106.

Wiles, A. D., Thompson, D. G. and Frantz, D. D., 2004. Accuracy assessment and interpretation for optical tracking systems. *Medical Imaging 2004*, International Society for Optics and Photonics, 421–432.

Wu, C. H., Sun, Y. N. and Chang, C. C., 2007. Three-dimensional modeling from endoscopic video using geometric constraints via feature positioning. *IEEE Transactions on Biomedical Engineering*, 54 (7), 1199–1211.

Wu, J.-R., Wang, M.-L., Liu, K.-C., Hu, M.-H. and Lee, P.-Y., 2014. Real-time advanced spinal surgery via visible patient model and augmented reality system. *Computer methods and programs in biomedicine*, 113 (3), 869–881.

Xie, J., Girshick, R. and Farhadi, A., 2016. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. B. Leibe, J. Matas, N. Sebe and M. Welling, eds., *Computer Vision – ECCV 2016*, Cham: Springer International Publishing, 842–857.

Xu, D., Ricci, E., Ouyang, W., Wang, X. and Sebe, N., 2017. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 161–169.

Yan, H., Zhang, S., Zhang, Y. and Zhang, L., 2018. Monocular depth estimation with guidance of surface normal map. *Neurocomputing*, 280, 86–100.

Yang, W., 2018. Context-aware computer aided inbetweening. *IEEE Transactions on Visualization and Computer Graphics*, 24 (2), 1049–1062.

Ye, M., Giannarou, S., Meining, A. and Yang, G.-Z., 2016. Online tracking and retargeting with applications to optical biopsy in gastrointestinal endoscopic examinations. *Medical Image Analysis*, 30, 14–157. URL `http://dx.doi.org/10.1016/j.media.2015.10.003`.

Ye, M., Johns, E., Handa, A., Zhang, L., Pratt, P. and Yang, G., 2017. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. *Proceedings 2017 Hamlyn Symposium on Medical Robotics*.

Yip, M. C., Lowe, D. G., Salcudean, S. E., Rohling, R. N. and Nguan, C. Y., 2012. Tissue tracking and registration for image-guided surgery. *Medical Imaging, IEEE Transactions on*, 31 (11), 2169–2182.

Yu, L., Efstathiou, K., Isenberg, P. and Isenberg, T., 2016. Cast: Effective and efficient user interaction for context-aware selection in 3d particle clouds. *IEEE Transactions on Visualization and Computer Graphics*, 22 (1), 886–895.

Zhang, D., Li, Z., Chen, K., Xiong, J., Zhang, X., Wang, L. et al., 2013. An optical tracker based robot registration and servoing method for ultrasound guided percutaneous renal access. *Biomed Eng Online*, 12, 47.

Zhang, L., Ye, M., Giannarou, S., Pratt, P. and Yang, G.-Z., 2017. Motion-compensated autonomous scanning for tumour localisation using intraoperative'ultrasound. *Medical Image Computing and Computer-Assisted Intervention ? MICCAI 2017*.

Zhang, R., Tsai, P.-S., Cryer, J. E. and Shah, M., 1999. Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21 (8), 690–706.

Zhang, X., Fronz, S. and Navab, N., 2002. Visual marker detection and decoding in ar systems: A comparative study. *Proceedings of the 1st International Symposium on Mixed and Augmented Reality*, IEEE Computer Society, 97.

Zhang, Z., 1998. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision*, 27 (2), 161–195. URL `https://doi.org/10.1023/A:1007941100561`.

Zhao, C., Sun, L. and Stolkin, R., 2017. A fully end-to-end deep learning approach

for real-time simultaneous 3d reconstruction and material recognition. *2017 18th International Conference on Advanced Robotics (ICAR)*, 75–82.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C. and Torr, P., 2015. Conditional random fields as recurrent neural networks. *International Conference on Computer Vision (ICCV)*.

Zhou, F., Duh, H. B.-L. and Billinghurst, M., 2008. Trends in augmented reality tracking, interaction and display: A review of ten years of ismar. *Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, IEEE Computer Society, 193–202.

Zhou, T., Brown, M., Snavely, N. and Lowe, D. G., 2017. Unsupervised learning of depth and ego-motion from video. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6612–6619.

Zhou, T., Tulsiani, S., Sun, W., Malik, J. and Efros, A. A., 2016. View synthesis by appearance flow. *European Conference on Computer Vision*.

Zhu, Q., Mai, J. and Shao, L., 2015. A fast single image haze removal algorithm using color attenuation prior. *IEEE Transactions on Image Processing*, 24 (11), 3522–3533.

Zoran, D., Isola, P., Krishnan, D. and Freeman, W. T., 2015. Learning ordinal relationships for mid-level vision. *2015 IEEE International Conference on Computer Vision (ICCV)*, 388–396.

Zou, Y. and Liu, P. X., 2017. A high-resolution model for soft tissue deformation based on point primitives. *Computer Methods and Programs in Biomedicine*, 148 (Supplement C), 113 – 121. URL `http://www.sciencedirect.com/science/article/pii/S0169260716306071`.