Two-stage deep regression enhanced depth estimation from a single RGB image

Jianyuan Sun, Zidong Wang, *Fellow, IEEE,* Hui Yu, *Senior Member, IEEE,* Shu Zhang, Junyu Dong*, Pengxiang Gao

Abstract—Depth estimation plays a significant role in industrial applications, e.g. augmented reality, robotic mapping and autonomous driving. Traditional approaches for capturing depth, such as laser or depth sensor based methods, are difficult to use in most scenarios due to the limitations of high system cost and limited operational conditions. As an inexpensive and convenient approach, using the computational models to estimate depth from a single RGB image offers a preferable way for the depth prediction. Although the design of computational models to estimate the depth map has been widely investigated, the majority of models suffers from low prediction accuracy due to the sole utilization of a one-stage regression strategy. Inspired by both theoretical and practical success of two-stage regression, we propose a two-stage deep regression model, which is composed of two state-of-the-art network architectures, i.e. the fully convolutional residual network (FCRN) and the conditional generation adversarial network (cGAN). FCRN has been proved to possess a strong prediction ability for depth prediction, but fine details in the depth map are still incomplete. Accordingly, we have improved the existing cGAN model to refine the FCRN-based depth prediction. The experimental results show that the proposed two-stage deep regression model outperforms existing state-of-the-art methods.

Index Terms—depth prediction, a single RGB image, the rough depth map, neural networks.

1 Introduction

Depth estimation plays an increasingly key role in industrial applications [1], [2]. In particular, depth estimation has widespread applications in the field of robotics [3], autonomous driving [4], augmented reality (AR) [5] and 3D modelling [6] etc. Due to the limitations of high system cost and limited operation availabilities, the traditional depth capturing approaches including the ones based on the laser and other depth sensors are facing unprecedented technical challenges and difficulties in most scenarios. On the contrary, using the computational models to estimate depth from a single RGB image offers a more feasible and preferable way to capture the depth map with a lower system cost and wider operational conditions. Designing computational models for accurately predicting depth information from a single RGB image is a challenging task, due to the inherent ambiguity of mapping the intensity or color measurement into a depth value. In particular, for the indoor scenes, there are plenty of geometric details and serious object occlusions, large texture and object structural variations, all of which conduces to the difficulty of accurate estimation depth. However, obtaining the availability and reasonably accurate depth information is a key element for many computer vision tasks and promoting the development of engineering [7], [8].

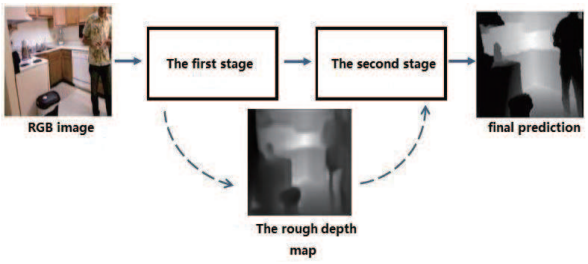


Fig. 1. **The framework of the proposed model**. In the first stage, the trained FCRN model is used to obtain a rough depth map. Moreover, a new cGAN model is used to refine the rough depth map and output the final prediction in the second stage.

Recently, Convolutional Neural Networks (CNNs) and their variants have been widely used to learn an inherent ambiguity relation between the RGB pixels and the depth information, which have achieved promising performance [9], [10], [11], [12], [13], [14], [15], [16]. Among the existing methods, most of the CNNs based approaches exploit some post-processing or regularization methods to refine the estimated depth map. For instance, CNNs are combined with the conditional random field (CRF) [12] to estimate the depth under the guidance of superpixel-wise depth information. Moreover, the multiple-scale CNNs

- *Jianyuan Sun was with the National Centre for Computer Animation, Faculty of Media and Communication, Bournemouth University, Bournemouth BH12 5BB, UK. (e-mail: sunj@bournemouth.ac.uk)*
  Shu Zhang and Junyu Dong were with the College of Information Science and Engineering, Ocean University of China, Qingdao 266100, China. (e-mail: zhangshu@ouc.edu.cn, dongjunyu@ouc.edu.cn).
  Junyu Dong also was with the Institute for Advanced Ocean Study, Qingdao 266100, China.
  Zidong Wang was with the Department of Computer Science, Brunel University, West London UB8 3PH, UK. (e-mail: zidong.Wang@brunel.ac.uk).
  Hui Yu was with the School of Creative Technologies, University of Portsmouth, Portsmouth PO1 2DJ, UK. (e-mail: hui.yu@port.ac.uk).
  Pengxiang Gao was with the School of Data Science and Software Engineering, Qingdao University, Qingdao, 266071, China. (e-mail: gaopengxiang@qdu.edu.cn).
- *Corresponding authors: Junyu Dong.*

models [9], [10], [15] have been proposed and developed to refine the predictions using a sequence convolution kernel of different scales to accurately capture the details of an object. Apart from these methods, some CNNs variants deal with the depth estimation tasks with an end-to-end learning architecture, which has achieved satisfying performance by increasing the layer number of the network architecture [14]. However, these methods not only need to hand-engineer the loss function for an acceptable result, but also require a large number of training data. Moreover, these methods usually have a higher complexity due to a large number of parameters involved in a complex or deep network architecture.

More importantly, the accuracy and reliability of the state-of-the-art RGB-based depth prediction deep network architectures [11], [14], [15] are still far from being practical. These methods are not effective in predicting the objects with far distance on NYU Depth v2 dataset [17] and Make3D dataset [18]. It is obvious that the existing methods only use the one-stage deep regression scheme to predict the depth from a single RGB image, normally only using a single deep learning model (CNNs or others) to predict depth. In fact, methods only using one-stage regression tend to lead to more inaccuracy for the depth prediction task due to the one-stage regression lacks the opportunity of re-learning or re-improving. In statistics [19], two-stage regression is proved to be able to provide better regression results than one-stage. The two-stage regression learning method has been successfully applied in the field of facial landmark detection [20] and other fields [21], [22], [23].

In this paper, inspired by both theoretical and practical success of two-stage regression, we propose a two-stage regression network, which is composed of two different types of state-of-the-art network architectures, i.e., the fully convolutional residual network (FCRN) and the conditional generation adversarial network (cGAN). The framework of the proposed model is shown in Fig. 1. The architecture of the FCRN consists of the fully convolutional architecture with residual learning. The convolution network has achieved great success in solving image tasks [24]. In particular, FCRN has been proved to possess a strong prediction ability for the depth prediction from a single RGB image [14]. However, the detailed information of the depth predictions is incomplete since the middle-level features are not fuse into the network. To achieve a more detailed depth map, we have improved the existing cGAN model [25] to tackle the regression task from the rough depth map to the ground truth depth map. In particular, the improved cGAN model allows us to use the maximum depth value to normalize the depth map for both indoor and outdoor scenarios, which can help improve the prediction accuracy for objects with far distance [26].

The contributions of this paper include:

- We propose a two-stage deep regression model for the task of depth estimation from a single RGB image according to the success of two-stage regression in both statistical theory and practice. The proposed model is composed of two different types of state-of-the-art network architectures. i.e. the fully convolutional residual network (FCRN) and the conditional generation adversarial network (cGAN). To our best knowledge, it is the first attempt to combine two different types of network architectures to tackle the depth estimation task.
- Inspired by the success of cGAN on regression tasks [26], [27], a new effectively refinement method is presented on the second stage of the proposed model. We improve the existing cGAN model with a "U-Net" generator to refine the FCRN-based depth prediction of the first stage. In particular, the proposed second stage model only needs a small number of training data compared with the state-of-the-art methods.
- With the experimental verification, in contrast to the existing methods using one-stage regression, the performance of the proposed two-stage regression model is superior to the previous fusion approaches on both indoor and outdoor scenarios [9], [10], [11], [12], [16].

2 Related Work

Depth prediction from a monocular RGB image has been receiving great attention, while it remains a very hard task due to the inherent ambiguity of mapping the RGB image intensity measure into the depth value. To tackle this task, many depth learning methods have been proposed to predict depth from a single RGB image in recent years, such as the convolutional neural networks (CNNs) and their variants. CNNs and its variants promote the state-of-the-art results for the challenging tasks of computer vision, which have also been applied to depth prediction from a single image and achieved great success [9], [10], [11], [12], [13], [14], [15], [16].

There exist two types of approaches to this problem: the multi-scale technique and the super-pixel pooling with conditional random field (CRF) algorithm. For the multi-scale technique, Eigen *et al.* [9] combined the depths from the global and refined network to obtain clear depth predictions. Their work later was extended to use a multi-scale convolutional network to predict depth by employing existing Alex and VGG networks [10]. Beyond that, most of the methods tended to combine the neural networks and a novel super-pixel pooling method of CRF to refine the depths. Liu *et al.* [11] first addressed this issue based on the fully convolutional networks and CRF. However, the prediction results are still far from the ground truth. Moreover, Wang *et al.* [13] and Li *et al.* [12] explored the benefit of the hierarchical CRF to refine the patch-wise predictions from the super-pixel level down to pixel level. Dan *et al.* [15] proposed a multi-scale continuous CRF as a sequential deep network for the depth prediction by exploring the side outputs of deep networks. In addition, Roy *et al.* [16] introduced an end-to-end architecture by integrating the random regression forests and convolutional neural networks to tackle the depth estimation. Recently, the residual network (ResNet) has successfully solved the gradient vanishing problem in the deeper networks, which have also been applied to depth prediction from a single image. For example, Laina *et*

al. [14] built a deeper fully convolutional residual networks and designed the up-sampling blocks to obtain a high-resolution depth. However, the detailed information of the depth predictions is incomplete due to the fact that the middle-level features are not fused into the network.

Besides, the utilization of convolutional neural networks, Generative Adversarial Networks (GANs) and its variants have also attracted many researchers' attention due to the GANs have achieved great success in many applications, such as image-to-image translation [25], face image generation [28], image in-painting [29], [30] and style transfer [31]. Moreover, there is some work using the GANs to tackle the depth prediction task. For example, Arun *et al.* [32] used GANs with flexible loss function to predict the depth map from a single RGB image on the KITTI dataset, which obtained depth prediction results more favorably compared with the state-of-the-art. In addition, Hyungjoo *et al.* [26] employed the existing cGAN model and design a fully convolutional multi-scale network to sequentially estimate the global and the local structures of the depth image. However, most of the depth prediction methods uses a single type of network model (CNNs or GANs) and enhances the network prediction ability by increasing the number of layers in the network. As a result, these networks require millions of training data for an acceptable estimation result.

In this paper, to reduce the complexity and improve the accuracy of depth prediction, we proposed a two-stage deep regression model that combines two different types of regression network architectures to track the depth prediction task, i.e., the trained FCRN [14] and the cGAN model. Our model combines the advantages of FCRN and cGAN, and uses the powerful prediction ability of cGAN on the regression task [26], [27] to further enhance the depth prediction. To our best knowledge, it is the first time to combine two different types of regression network models on the task of depth prediction.

3 Proposed Method

The proposed method aims to predict the depth map from a single RGB image. We specifically design a new two-stage deep regression model to accurately predict the depth image from a single RGB image. In the first stage, the initial (rough) depth map is obtained using a pre-trained fully convolutional residual network model (FCRN) [14]. In the second stage, a new conditional generative adversarial network (cGAN) [25] is proposed to refine the FCRN-based depth prediction.

3.1 Two-stage deep regression model

Inspired by the success of two-stage regression in both theory and practice, we propose a two-stage deep regression model to predict the depth map from a single RGB image. The proposed model is composed of two different network architectures. i.e. the fully convolutional residual network (FCRN) and the improved conditional generation adversarial network (cGAN).

The network architecture of the proposed model is illustrated in Fig. 2. For the first stage model, the fully convolutional residual network (FCRN) is first proposed by Laina *et al.* [14]. In particular, FCRN model has achieved promising results in the task of the depth prediction. The FCRN architecture is based on the ResNet-50 and uses a new up-projection block, which yields an output of roughly half the input resolution [14]. The input size of FCRN is 483×483. To reduce billions of parameters and dozens of GB memory generated by the full convolution network, FCRN use a new up-sampling blocks that contain fewer weights. Moreover, the ResNet-50 [33] can make the FCRN muck deeper and prevent gradients from vanishing or degradation. Therefore, there are large receptive fields for the FCRN with the deep architecture.

Using the up-projection block is the key technology to make the FCRN achieve good prediction results. The up-projection block is a new up-sampling block and extends the idea of the projection connection [33] to up-convolutions. Here, the up-convolutions are the up-sampling res-blocks. The main idea is to introduce a simple of 3×3 or 5×5 convolution after the up-convolution to add a projection connection from the lower resolution feature map to the result. Beyond that, Laina *et al.* developed the chain up-projection blocks in FCRN that allowed high-level information to be more efficiently passed forward in the network while progressively increasing feature map sizes [14].

For the FCRN optimization, it uses a reverse Huber [34], [35] as the loss function B.

$$B(x) = \begin{cases} |x| & |x| \leq c, \\ \frac{x^2+c^2}{2c} & |x| > c. \end{cases} \tag{1}$$

As shown in Eq. (1), the Berhu loss is equal to the $\mathcal{L}_1(x) = |x|$ norm when $x \in [-c, c]$, and equal to $\mathcal{L}_2$ norm when $|x| > c$. Here, $c = \frac{1}{5} max_i(|\tilde{y}_i - y_i|)$, where i represents the index of each pixel in each image in the current batch.

In the proposed model, we first obtain the initial depth map from a single RGB image by using the trained FCRN model [14] in the first stage. The output resolution of FCRN is 160×128. To carry out the prediction of the second stage, we use a bilinear interpolation method to up-sample the obtained depth maps back to the size of 256×256 due to the input size of the improved cGAN model is 256×256.

For the second stage, the improved cGAN model is used to refine the depth predictions of the first stage FCRN model. Inspired by the work of image-to-image translation [25], we improved the original cGAN model [25]to tackle the rough depth map d to the ground truth depth map g translation task. Like the original cGAN model, the improved cGAN model also consists of a generator network $G(\cdot)$ and a discriminator network $D(\cdot)$.

For the generator $G(\cdot)$ of the improved cGAN, it has the skip connections to shuttle the low-level information directly across the net. Moreover, the generator follows the shape of "U-NET" [36] like the generator in the model of image-to-image translation [25], as shown in Fig. 3. Specifically, there are skip connections between each layer i and layer $n - i$, where n indicates the total number of layers. In order to adapt the task of refining the rough depth maps, we have changed the number of spatial filters at each convolution layer and de-convolution layer. In

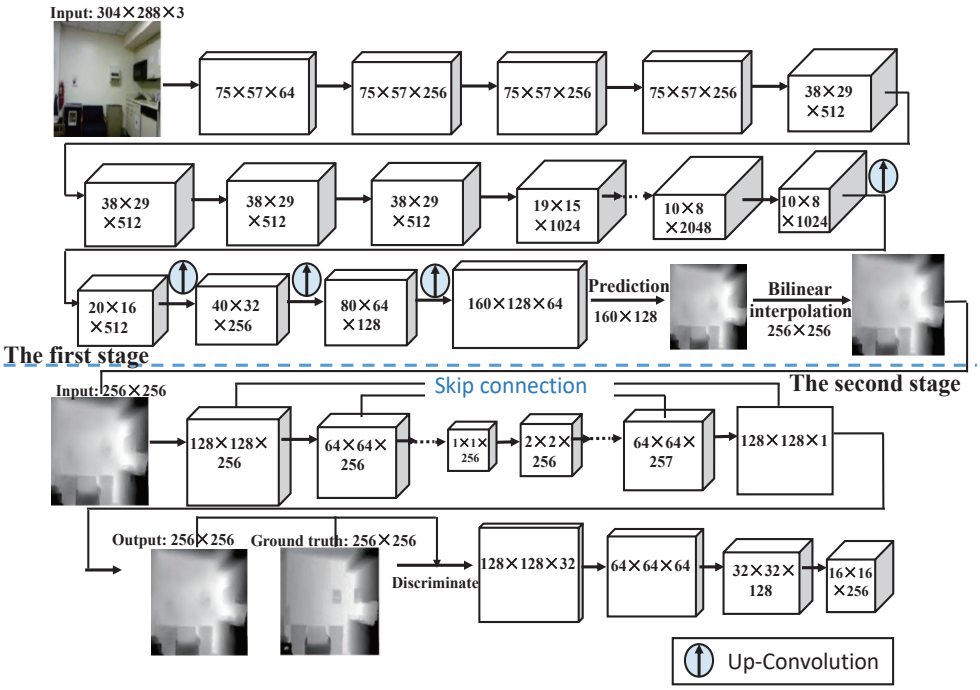


Fig. 2. The two-stage deep regression architecture. In the first stage, a trained FCRN is used to obtain the rough depth maps [14]. The FCRN builds upon ResNet-50, and uses a novel up-sampling blocks. In the second stage, we propose a new cGAN based on the existing cGAN model [25] to tackle the task of refining the FCRN-based depth prediction. Note that, the new cGAN model adopts the "U-Net" for the generator.

addition, the improved cGAN uses 3×3 spatial filters with stride 2 for all convolution layer instead of the original cGAN uses 5×5 spatial filters with stride 2 in the image-to-image translation task. Specially, we only use the dropout in the first three convolution layers of "U-Net" decoder.

For the discriminator, to learn the high-frequency information from the depth map, we restrict the discriminator to use the structure of the local image patches. That is, the discriminator of the improved only penalizes the scale of patches for the depth maps, which tried to classify each $N \times N$ patch in a depth map being real or fake. Inspired by the structure of existing model [25], using the convolution operations on discriminator to achieve the scale discrimination of the depth map. Moreover, we run this discriminator convolutional across the depth map and average all responses to provide the ultimate output of discriminator D. In particular, we use 70×70 discriminator patch according to the experience.

For the improved cGAN optimization, like the original cGAN model, it alternatively optimizes $D(\cdot)$ along with a generator $G(\cdot)$ to solve the follow min-max optimization problem [25], [37]:

$$\min_G \max_D \mathbb{E}_{g\sim p_G}[log(1 - D(g, G(g))] + \mathbb{E}_{g',d'\sim p_g t} [logD(g^{'}, d^{'})] + \lambda \mathbb{E}_{g\sim p_G, d\sim p_{gt}}[\|G(g) - d\|_1], \quad (2)$$

where the discriminator $D(\cdot)$ is trained to distinguish samples from the ground truth distribution p_{gt} and the generative distribution p_G. (g, d) and $(g^{'}, d^{'})$ are sampled from the rough depth map to the ground truth depth map, and λ represents the relative weighting factor.

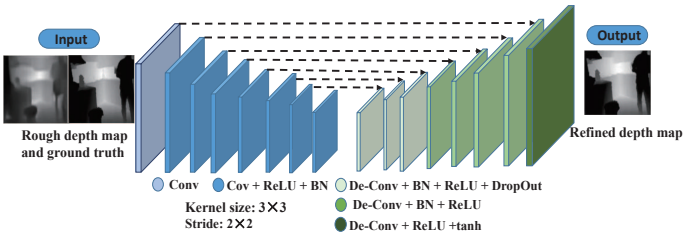


Fig. 3. **Generator architecture of the improved cGAN**. The "U-net" [36] based encoder-decoder with skip connections is similar to the generator of the original cGAN [25].

For the optimization of the refine depth map, Eq. (2) consists of an adversarial loss (the first two terms) and a pixel-wise reconstruction loss (the last term). The discriminator D takes the generated depth map from the generator G and the ground-truth depth map as the inputs and discriminates whether these are network output or not. That is, Eq. (2) can enable one to train the generator network G and deceive discriminator network D. Accordingly, we can obtain the results that are highly close to the ground truth depth map or are indistinguishable by D.

4 Experiments

To demonstrate the effectiveness of the proposed model for monocular depth prediction, we carry out the experiments on two public datasets including the indoor dataset (NYU Depth v2) and the outdoor dataset (Make3D). For the

quantitative evaluation of the results, we use some metrics from the existing works:

- Abs rel: $\frac{1}{N}\sum_{y_i \in |N|} \frac{|y_i - y_i^*|}{y_i^*}$
- Rms: $\sqrt{\frac{1}{N}\sum_{y_i \in |N|} |y_i - y_i^*|^2}$
- Average $\log_{10}$ error: $\log_{10} = \frac{1}{N}\sum_{y_i \in |N|} |\log_{10}(y_i) - \log_{10}(y_i^*)|$
- Threshold: $\max(\frac{y_i^*}{y_i}, \frac{y_i}{y_i^*}) = \delta < threshold(t \in [1.25, 1.25^2, 1.25^3])$

where y_i denotes the estimated depth, y_i^* is the corresponding ground truth depth, N is the total number of pixels.

Implementation Details For the proposed two-stage model, we first use the trained FCRN model [14] to obtain the rough depth map. Then, we combine each estimated rough depth map and the corresponding ground truth depth map in a side by side manner as the input of the improved cGAN model. Moreover, we introduce the details of the proposed cGAN model for refining the FCRN-based depth prediction on the NYU Depth v2 dataset and Make3D dataset respectively. To explicitly describe the details of the improved cGAN architecture, we define Ck to represent a Convolution-BatchNorm-ReLU layer with k spatial filters and CDk represents a Convolution-BatchNorm-Dropout-ReLU layer with the dropout rate of 50%.

For the generator architecture in the improved cGAN model, the encoder and decoder architecture of "U-Net" are as follows.

For the NYU Depth v2 dataset, "U-Net" encoder is

$$C256-C256-C512-C512-C512-C256-C256-C256,$$

and "U-Net" decoder is

$$CD512-CD384-C576-C513-C513-C257-C257.$$

For the Make3D dataset, "U-Net" encoder is

$$C256-C512-C512-C512-C512-C256-C256-C256,$$

and "U-Net" decoder is

$$CD384-CD320-C578-C513-C513-C513-C257.$$

For the discriminator architecture, the 70×70 discriminator architecture for the NYU Depth v2 dataset is

$$C32-C64-C128-C256,$$

and the 70×70 discriminator architecture for the Make3D dataset is

$$C1-C1-C1-C1-C1.$$

Note that, all ReLUs are leaky, with slope 0.2 in the improved cGAN model. In addition, we use TensorFlow [38] deep learning framework to implement the proposed model for the validation experiments. The improved cGAN model is trained on a single NVIDIA GeForce GTX TITAN with $12GB$ memory. The evaluation results on the NYU Depth v2 dataset and Make3D dataset are discussed in the following subsections.

4.1 NYU Depth v2 Dataset

The NYU Depth v2 Dataset is a publicly available dataset, which contains 645 indoor scenes. To evaluate the performance of the proposed model, we follow the official training and testing assignment, i.e., the training dataset consists 249 scenes with 795 images, and testing dataset consists 215 scenes with 654 images. For the proposed model, the training process of the first stage FCRN model is the same as that of Laina et al [14]. That is, nearly 95k pairs of RGB-D images of the NYU Depth v2 are obtained by sampling equally-spaced frames for each training sequence and using the data augment methods to train the first stage FCRN model. In particular, we down-sample the original images of size 640×480 pixels to $1/2$ resolution and center-crop to 304×228 pixels, as input to the FCRN. Moreover, the FCRN is trained with a batch size of 16 for approximately 20 epochs. The starting learning rate is set to 10^{-2} for all layers, which is gradually reduced every $6-8$ epochs. The momentum is 0.9.

The trained FCRN model [14] is used as the first stage FCRN model to obtain the roughly estimated depth map for the 795 training dataset and 654 testing dataset, respectively. Then, we combine each roughly estimated depth map and the corresponding ground truth depth map from 795 training dataset and 654 testing dataset in a side by side way before implementing the second stage of the model. In this way, it can form up the new training dataset and test dataset for the second stage of the model. That is, we use the proposed cGAN model to refine the rough depth maps based on the corresponding ground truth. Here, the size of the FCRN-based rough depth map is 160×128, and the size of the input depth map of cGAN is 256×256. Therefore, the estimated rough depth maps are up-sampled to 256×256 using bilinear interpolation before obtaining the new training and testing dataset. In particular, we only use the 795 training data to train the cGAN model in the second stage for the NYU Depth v2 dataset.

For the parameters of the second stage model, the proposed cGAN model is trained with a batch size of 1 for about 20 epochs. Following [25], we use minibatch SGD and Adam solve [39] with the learning rate being 0.0002, and the momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. To compare our results with state-of-the-art methods, we up-sample the estimated depth maps back to the original size 640×480. The quantitative comparisons on NYU Depth v2 are shown in Table 1. From the comparison results, our model achieves the promising results.

Moreover, the visualization results of the estimated depth map are shown in Fig. 4. We compare a CNN-based method with the multi-scale architecture [10], a method of fusing the CNNs and the CRF [11], the single FCRN method [14] and the single cGAN model [25] to show the performance among different models. From Fig. 4, it is obvious that our results are closer to the realistic.

To illustrate the effectiveness of the proposed model for predicting the depth value of objects with far distance. The visualization results of the details of the estimated depth map are presented in Fig. 5. Specially, we compare a depth prediction method of fusing the CNNs and the classical

random forests regression algorithm [16] with the proposed model. As show in Fig. 5, our model is more effective for predicting the depth value of objects with far distance.

TABLE 1
Comparison results of Monocular depth prediction on the NYU Depth v2 dataset [17]. The values are those reported original reported from the authors in their respective paper.

NYU Depth v2	rel	rms	log_{10}	δ_1	δ_2	δ_3
Karsh *et al.* [40]	0.374	1.12	0.134	–	–	–
Ladicky *et al.* [41]	–	–	–	0.542	0.829	0.941
Liu *et al.* [42]	0.335	1.06	0.127	–	–	–
Li *et al.* [12]	0.232	0.821	0.094	0.621	0.886	0.968
Liu *et al.* [11]	0.230	0.824	0.095	0.614	0.883	0.971
Wang *et al.* [13]	0.220	0.745	0.094	0.605	0.890	0.970
Eigen *et al.* [9]	0.215	0.907	–	0.611	0.887	0.971
Roy and Todorovic [16]	0.187	0.744	0.078	–	–	–
Eigen and Fergus [10]	0.158	0.641	–	0.769	0.950	0.988
Lei *et al.* [43]	0.151	0.572	0.064	0.787	0.948	0.986
The single **FCRN** [26]	0.127	0.573	0.055	0.811	0.953	0.988
The single **cGAN**	0.184	1.573	–	0.186	0.348	0.489
Our model	**0.114**	**0.563**	**0.049**	**0.812**	**0.955**	**0.989**

TABLE 2
Comparison results of Monocular depth prediction on the Make3D dataset [18].

Make3D	rel	rms	log_{10}
Karsh *et al.* [40]	0.355	9.20	0.127
Liu *et al.* [42]	0.335	9.49	0.137
Liu *et al.* [11]	0.314	8.60	0.119
Li *et al.* [12]	0.278	7.19	0.092
Lei *et al.* [43]	0.207	6.90	0.084
The single **FCRN** [26]	0.175	4.45	0.072
The single **cGAN**	0.336	8.38	0.187
Our model	**0.167**	**4.32**	**0.064**

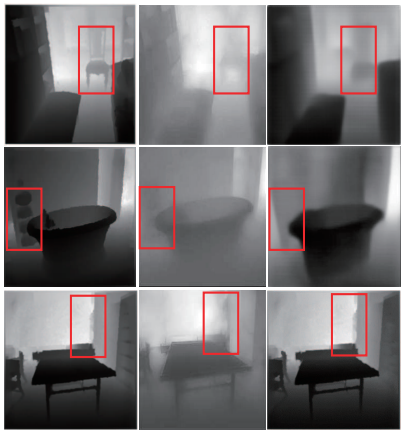


Fig. 5. **Details of depth prediction on the NYU Depth v2**. From left to right: ground truth, the results of Roy and Todorovic [16] and our results.

4.2 Make3D Dataset

The Make3D dataset is a publicly available dataset for outdoor scenarios, which consists of 400 training and 134 testing images. For the proposed model, the training process of the first stage FCRN model is the consistent with

Fig. 6. **Qualitative results on the Make3D**. For our results, pixels that distances $> 70m$ are masked out according to the existing work [14], [42].

that of Laina et al [14]. That is, nearly 15k pairs of RGB-D images of the Make3D are obtained by using the data augment methods to train the first stage FCRN model. In particular, we down-sample the original images of size 345×460 pixels to $1/2$ resolution and center-crop to 304×228 pixels, as input to the FCRN. Moreover, the FCRN is trained with a batch size of 16 for approximately 30 epochs. The starting learning rate is set to 0.01 for all layers, which is gradually reduced by optimizing the loss function. The momentum is 0.9.

The trained FCRN model [14] is used as our first stage FCRN model to obtain the roughly estimated depth map for the 795 training dataset and 654 testing dataset, respectively. Then, we combine each roughly estimated depth map and the corresponding ground truth depth map for the 400 training and 134 testing images in a side by side way before implementing the second stage of the model. Then, we can obtain the new training and testing dataset. That is, we use the proposed cGAN model to refine the rough depth maps based on this new training and testing dataset. Moreover, we expand the new training dataset from 400 to 15598 by using some methods of the training data augmentation [44]. The dataset augmentation methods include the scale transformation, i.e., the roughly estimated depth maps are scaled by a random number $s \in [10, 15]$ and the corresponding ground truth are divided by s; flips, i.e., the rough estimated depth maps and the corresponding ground truth maps are both horizontally flipped with a 90% chance; adjust the brightness of the rough estimated depth maps and the corresponding ground truth depth maps.

Fig. 4. **Qualitative results on the NYU Depth v2**. From left to right: RGB image, ground truth, the results of Liu *et al.* [11], the results of Eigen and Fergus [10], the results using FCRN [26] and our results.

For the parameters of the second stage model, the proposed cGAN model is trained with a batch size of 1 for about 20 epochs. We also use minibatch SGD and Adam solve [39] with the learning rate 0.0002, and the momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. To compare our results with state-of-the-art, we up-sample the estimated depth maps back to 345×460. Moreover, the quantitative evaluation of our model is shown in Table 2. In particular, considering the low-resolution ground truth and the inaccuracy over the range $70m$ (e.g. sky pixels mapping is 80m), we train against ground truth depth maps and obtain the quantitative results by masking out pixels of distances over $70m$ according to the existing work [14], [42]. From the results, we can see that our model represents the state-of-the-art. In addition, the qualitative results on Make3D are shown in Fig. 6, from which the predictions of our method are closer to the realistic.

4.3 Result Analysis

Based on the above experiments, the performance of the proposed two-stage model is better than that of other state-of-the-art methods on the depth prediction task. The reasons for this result are as follows.

As we all know, the two-stage regression model has one more regression process than the one-stage regression model. That is to say, our model has one more process of correcting prediction error than the existing one-stage regression model. In the first stage of our model, the trained FCRN obtains the original depth maps from the single RGB image. This regression process is equivalent to adding a probabilistic priori to the training dataset in the second stage. The first step regression will reduce the computational search space of the second stage regression and improve the prediction accuracy. Accordingly, in the second stage, the improved cGAN model further improves the accuracy of FCRN-based predictions. In addition, from the Table 1 and Table 2, we found that the cGAN model alone cannot accurately predict the depth image from a single RGB image. In fact, the cGAN model can achieve high-precision conversion for the two images with similar styles [25]. Depth prediction from a single RGB image is a very hard task due to the inherent ambiguity of mapping the RGB image intensity measure into the depth value. Therefore, we cannot obtain a depth map with higher accuracy by using the cGAN model alone.

In most cases, the multi-stage learning approach yields better predictive results than the one-stage learning approach. Therefore, we have tried to add the third stage regression model to the proposed model. i.e. FCRN-cGAN-cGAN. The quantitative results on NYU Depth v2 and Make3D are shown in Table 3and Table 4. It is not difficult to find that the prediction results are worse than those of the two-stage regression model. The three-stage regression model did not significantly improve the prediction results of the two-stage regression model. In fact, for some models, the two-stage regression model is sufficient to solve practical problems, such as the Two-Stage Least Squares (2SLS) Regression Analysis [19], [45].

TABLE 3
Comparison results between the proposed two-stage model and three-stage model on the NYU Depth v2 dataset [17].

NYU Depth v2	rel	rms	log_{10}	δ_1	δ_2	δ_3
Two-stage model	**0.114**	**0.563**	**0.049**	**0.812**	**0.955**	**0.989**
Three-stage model	0.124	1.150	0.070	0.627	0.836	0.926

TABLE 4
Comparison results between the proposed two-stage model and three-stage model on the Make3D dataset [18].

Make3D	rel	rms	log_{10}
Two-stage model	**0.167**	**4.32**	**0.064**
Three-stage model	0.173	4.56	0.082

5 Conclusion and Future Research

In this work, we propose a two-stage regression model to tackle the task of depth estimation from a single RGB image. Unlike most existing methods, which use the only one-stage regression method, this paper is the first attempt to explore the two-stage regression method for predicting

the depth from a single RGB image. The two-stage regression model is proved to be able to provide a better performance than its one-stage counterpart. Previous one-stage regression methods usually require a lot of training data due to they mainly increase the number of layers and the complexity of the model to achieve satisfactory results. In the second stage of our model, only a few training data sets are needed to improve the results of the existing depth prediction model. Our model also has some disadvantages. i.e., the proposed two-stage model has more parameters and complexity than the existing one-stage methods.

The success of our model lies in combines two different types of state-of-the-art network architectures. i.e. the fully convolutional residual network (FCRN) and the conditional generation adversarial network (cGAN). Based on the research results, we can conclude that the proposed second stage model can not only improve the accuracy of FCRN-based depth prediction but also improve the depth prediction results of existing depth models. In the future, we will use the proposed second stage model to improve the results of other existing models for solving different tasks.

ACKNOWLEDGMENT

This work was supposed by the EPSRC through project 4D Facial Sensing and Modeling (EP/N025849/1), the National Natural Science Foundation of China (NSFC)(No. 61271405, No. 41906177, No. U1706218, No. 41927805), the Fundamental Research Funds for the Central Universities 201964022 and the National Key R&D Program of China (Grant No. 2018AAA0100602), the Shandong Provincial Natural Science Foundation, china (No. ZR2018ZB0852), Marine Data Science strategic project (No. L1824025) and the International Science and Technology Cooperation Program of China (ISTCP) (No. 2104DFA10410).

REFERENCES

[1] A. Letouzey, B. Petit, and E. Boyer, "Scene flow from depth and color images," in *British Machine Vision Conference*, pp. 1–11, 2011.

[2] M. Nielsen, D. C. Slaughter, and C. Gliever, "Vision-based 3d peach tree reconstruction for automated blossom thinning," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 188–196, 2012.

[3] K. Withanage, I. Lee, R. Brinkworth, S. Mackintosh, and D. Thewlis, "Fall recovery sub-activity recognition with rgb-d cameras," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 6, pp. 2312–2320, 2016.

[4] T. H. S. Li, S. J. Chang, and Y. X. Chen, "Implementation of human-like driving skills by autonomous fuzzy behavior control on an fpga-based car-like mobile robot," *IEEE Transactions on Industrial Electronics*, vol. 50, no. 5, pp. 867–880, 2003.

[5] J. M. Andujar, A. Mejias, and M. A. Marquez, "Augmented reality for the improvement of remote laboratories: An augmented remote laboratory," *IEEE Transactions on Education*, vol. 54, no. 3, pp. 492–500, 2011.

[6] X. Xu, A. Song, D. Ni, H. Li, P. Xiong, and C. Zhu, "Visual-haptic aid teleoperation based on 3d environment modeling and updating," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 10, pp. 6419–6428, 2016.

[7] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision (ECCV)*, pp. 746–760, 2012.

[8] J. B. A. E. Raia Hadsell, Pierre Sermanet and M. Scoffier, "Learning long-range vision for autonomous off-road driving," *Journal of Field Robotics*, pp. 120–144, 2009.

[9] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *International Conference on Neural Information Processing Systems*, pp. 2366–2374, 2014.

[10] E. David and F. Rob, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *IEEE International Conference on Computer Vision*, pp. 2650–2658, 2015.

[11] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," pp. 5162–5170, 2014.

[12] B. Li, C. Shen, Y. Dai, A. V. D. Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1119–1127, 2015.

[13] W. Peng, X. Shen, Z. Lin, and S. Cohen, "Towards unified depth and semantic prediction from a single image," in *Computer Vision and Pattern Recognition*, pp. 2800–2809, 2015.

[14] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," pp. 239–248, 2016.

[15] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," pp. 161–169, 2017.

[16] A. Roy and S. Todorovic, "Monocular depth estimation using neural regression forest," in *Computer Vision and Pattern Recognition*, pp. 5506–5514, 2016.

[17] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*, pp. 746–760, 2012.

[18] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824–840, 2009.

[19] D. S. Moore, "Statistics," *W.H.Freeman and Co Ltd*, 1979.

[20] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou, "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3691–3700, 2017.

[21] C. Y. Yeh, C. W. Huang, and S. J. Lee, "A multiple-kernel support vector regression approach for stock market price forecasting," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2177–2186, 2011.

[22] Y. Li, J. H. Gilmore, J. Wang, M. Styner, W. Lin, and H. Zhu, "Two-stage multiscale adaptive regression methods for twin neuroimaging data," *IEEE Transactions on Medical Imaging*, vol. 31, no. 5, pp. 1100–1112, 2012.

[23] N. Cummins, V. Sethu, J. Epps, J. R. Williamson, T. F. Quatieri, and J. Krajewski, "Generalized two-stage rank regression framework for depression score prediction from speech," *IEEE Transactions on Affective Computing*, no. 99, pp. 1–1, 2018.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1106–1114, 2012.

[25] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," pp. 5967–5976, 2016.

[26] H. Jung, Y. Kim, D. Min, C. Oh, and K. Sohn, "Depth prediction from a single image with conditional adversarial networks," in *IEEE International Conference on Image Processing*, pp. 1717–1721, 2017.

[27] T. Li, X. Liu, and S. Su, "Semi-supervised text regression with conditional generative adversarial networks," in *IEEE International Conference on Big Data*, pp. 5375–5377, 2018.

[28] J. Choe, P. Song, K. Kim, J. H. Park, D. Kim, and H. Shim, "Face generation for low-shot learning using generative adversarial networks," in *IEEE International Conference on Computer Vision Workshop*, pp. 1940–1948, 2017.

[29] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Computer Vision and Pattern Recognition*, pp. 2536–2544, 2016.

[30] T. F. Y. Vicente, L. Hou, C. P. Yu, M. Hoai, and D. Samaras, "Large-scale training of shadow detectors with noisily-annotated shadow examples," in *European Conference on Computer Vision*, pp. 816–832, 2016.

[31] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *European Conference on Computer Vision*, pp. 702–716, 2016.

[32] A. CS Kumar, S. M. Bhandarkar, and M. Prasad, "Monocular depth prediction using generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 300–308, 2018.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[34] L. Zwald and S. Lambertlacroix, "The berhu penalty and the grouped effect," *Statistics*, 2012.

[35] A. B. Owen, "A robust hybrid of lasso and ridge regression," *Contemporary Mathematics*, vol. 43, pp. 59–72, 2007.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

[37] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *International Conference on Neural Information Processing Systems (NIPS)*, pp. 2672–2680, 2014.

[38] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv: Distributed, Parallel, and Cluster Computing*, 2016.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ICLR*, 2015.

[40] K. Karsch, C. Liu, and S. B. Kang, "Depth extraction from video using non-parametric sampling," in *European Conference on Computer Vision*, pp. 775–788, 2012.

[41] J. Shi and M. Pollefeys, "Pulling things out of perspective," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 89–96, 2014.

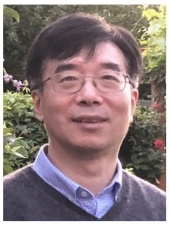
[42] M. Liu, M. Salzmann, and X. He, "Discrete-continuous depth estimation from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 716–723, 2014.

[43] L. He, G. Wang, and Z. Hu, "Learning depth from single images with deep neural network embedding focal length," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 27, no. 9, pp. 4676–4689, 2018.

[44] F. Mal and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," *international conference on robotics and automation*, pp. 1–8, 2018.

[45] K. A. Bollen, "An alternative two stage least squares (2sls) estimator for latent variable equations," *Psychometrika*, vol. 61, no. 1, pp. 109–121, 1996.

Jianyuan Sun received her Ph.D. degree in Computer Application Technology in June 2019, from the College of Information Science and Engineering at Ocean University of China. She is currently a research associate at the National Centre for Computer Animation, Bournemouth University, Poole, UK. She research interests include deep learning, machine learning, computer vision and 3D reconstruction.

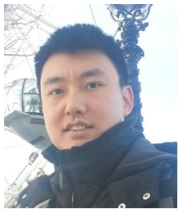


Wang Zidong (SM'03-F'14) was born in Jiangsu, China, in 1966. He received the B.Sc. degree in mathematics in 1986 from Suzhou University, Suzhou, China, and the M.Sc. degree in applied mathematics in 1990 and the Ph.D. degree in electrical engineering in 1994, both from Nanjing University of Science and Technology, Nanjing, China.

He is currently Professor of Dynamical Systems and Computing in the Department of Information Systems and Computing, Brunel University London, U.K. From 1990 to 2002, he held teaching and research appointments in universities in China, Germany and the UK. Prof. Wang's research interests include dynamical systems, signal processing, bioinformatics, control theory and applications. He has published more than 300 papers in refereed international journals. He is a holder of the Alexander von Humboldt Research Fellowship of Germany, the JSPS Research Fellowship of Japan, William Mong Visiting Research Fellowship of Hong Kong.

Prof. Wang serves (or has served) as the Editor-in-Chief for Neurocomputing, the Deputy Editor-in-Chief for International Journal of Systems Science, and an Associate Editor for 12 international journals, including IEEE Transactions on Automatic Control, IEEE Transactions on Control Systems Technology, IEEE Transactions on Neural Networks, IEEE Transactions on Signal Processing, and IEEE Transactions on Systems, Man, and Cybernetics - Part C. He is a Fellow of the IEEE, a Fellow of the Royal Statistical Society and a member of program committee for many international conferences.

Hui Yu is a Professor with the University of Portsmouth, UK. His research interests include methods and practical development in visual computing, machine learning and AI with the applications focusing on human-machine interaction, multimedia, virtual/augmented reality and robotics as well as 4D facial expression generation, perception and analysis. He serves as an Associate Editor for IEEE Transactions on Human-Machine Systems and Neurocomputing journal.



Shu Zhang received his PhD degree in Computer Application Technologies from Ocean University of China, Qingdao, China. He was a research associate at University of Portsmouth, Portsmouth, UK. He is currently a lecturer with Ocean University of China, Qingdao, China. His main research interests include image processing, feature matching, 3D reconstruction, underwater image analysis.

Junyu Dong received the Ph.D. degree in 2003 from Heriot-Watt University, UK. Junyu Dong joined Ocean University of China in 2004. From 2004 to 2010, Dr. Junyu Dong was an associate professor at the Department of Computer Science and Technology. He became a Professor in 2010 and is currently the Head of the Department of Computer Science and Technology. Currently, Prof. Dong is the Chairman of Qingdao Young Computer Science and Engineering Forum (YOCSEF Qingdao). He is a member of ACM and IEEE. Prof. Dong¡¯s research interests include texture perception and analysis, 3D reconstruction, video analysis and underwater image processing.

Pengxiang Gao received his BSc and MSc from the Northeast University in 1985 and 1989 respectively. He joined Qingdao University in 2004 and he is currently a professor and the Head of the School of Data Science and Software Engineering. His research interest includes pattern recognition, big data and robot technology.