Improved Attentive Pairwise Interaction (API-Net) for Fine-Grained Image Classification

Ong Zu Yet[1], Taha H. Rassem[2], Md Arafatur Rahman[3], and M. M. Rahman[4]

[1]Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang, Pekan, Malaysia
Email: zu.y.ong528@gmail.com

[2]Faculty of Science and Technology, Bournemouth University, Poole BH12 5BB, UK
Email: tahahussein@ieee.org

[3]University of Wolverhampton, School of Mathematics and Computer Science, Wolverhampton, WV1 1LY, UK

[4]Department of Mechanical Engineering, College of Engineering, Universiti Malaysia Pahang, 26300 Kuantan, Malaysia.
Email: mustafizur@ump.edu.my

***Abstract*—Fine-grained classification is a challenging problem as one has to deal with a similar class of objects but with various types of variations. For more elaboration, they are almost similar and have subtle differences, and are confusing. In this study, aircraft will be the fine-grained object to be focused on. Aircraft which has almost similar shapes and patterns can be hardly recognized even for humans, especially those who haven not gone through any training. In recent years, a lot of proposed methods addressed to solve the difficulties in fine-grained problems by learning contrastive clues from an image. This study aims to increase the accuracy of the Attentive Pairwise Interaction Network (API-Net) by introducing data augmentation into the network structure. Some of the previous studies proved that data augmentation does help improve a network. So, this study is going to modify the API-Net with different data augmentation settings. In this study, various settings have been introduced to the API-Net. Several experiments had been done with a simple modification where a portion of the train dataset's images will randomly convert into greyscale images. These settings are, only brightness & contrast 0.2, only grayscale 0.3, only grayscale 0.5, brightness & contrast 0.2 with grayscale 0.3, and brightness & contrast 0.2 with grayscale 0.5. As a result, the proposed modification achieved with 92.74% with brightness & contrast 0.2, 92.80% on brightness & contrast 0.2 with grayscale 0.5, and 92.86% on brightness & contrast 0.2 with grayscale 0.3. While grayscale 0.3 alone achieve 93.25% and grayscale 0.5 alone achieve 93.46% compared with the original results which reached 92.77%.**

I. Introduction

Computer vision supports a wide range of applications such as Optical Character Recognition(OCR) [1], object detection and recognition [2], vision biometrics [3], medical imaging, smart city, smart transport, social media, etc. OCR converts displayed or printed text or handwritten text into digital form. Still, there is a lot of challenges found in OCR because of the existence of different languages, fonts, style, complex rules of languages, and handwriting [1]. Vision biometrics deals with the physical and behavioral characteristics of an individual. Physical characteristics such as face, iris, and fingerprints, and behavioral characteristics such as signature move and gait. Challenges in vision biometrics such as a large number of identities, intrapersonal variation, extracting biometric information have to be overcome to make the system effective [3]. Object detection and recognition have been widely applied in multiple works. These includes generic object detection [4], [5], [6], [7], [8], specific object detection such as road traffic sign detection [9], pedestrian detection [10], [11], vehicle detection [12] and face detection [13]. While these also extended their application into the medical field such as in [14], [15], [16].

Recognizing an object in the image has been a longstanding and challenging problem in computer vision. For several decades, the interest of research in this area has been actively involved. The goal of this approach is to detect the object in the image, determine and classify it to the correct category, class, etc. For decades, the emergence of deep learning in computer vision has revolutionized to be able to solve tasks such as image classification, image detection, image segmentation, etc. Deep learning algorithms such as Convolutional Neural Networks (CNNs) utilize the advantage of properties such as local connections, shared weights, pooling, loads of layers, and composition hierarchies [17] made significant advancements in object recognition. Several deep learning including CNNs on object detection and object recognition are summarized in [2].

Since the remarkable breakthrough accomplishment of CNN makes possible in visual recognition by some convolutional network such as ResNet [6], DenseNet [18], image recognition achieves higher accuracy than before, but, these model often have limited capabilities in recognizing fine-grained categorical images due to their highly confusing and high similarity characteristics. Fine-grained visual classification (FGVC) will be the focus of this study. The main objective in fine-grained recognition is to aim for the subtle discriminative details of subordinate categories within a basic level category and differentiate them. In another word, fine-grained image classification deals with the objects of more intra-class variance

978-1-6654-3305-1/21/$31.00 ©2021

than inter-class variance, and these images don't have much noticeable difference and almost look alike. Therefore, a lot of works have to be done on top of these CNN to make the fine-grained image classification even better. Several fine-grained frameworks have been proposed in the past by finding prominent regions [19], [20]. A more recent proposed work such as Yang et al. [21] uses a Navigator-Teacher-Scrutinizer system as multi-agent cooperation arranged in multi-stages. In an effort from Sun et al. [22] proposed a way to explicitly force the network to find the subtle differences among closely related class with a diversification block and a gradient boosting loss. Some of the latest work such as Attention Pairwise Interaction Network (API-Net) from Zhuang et al. [23] focus on making the machine learning from not just one image, but two. API-Net adaptively discovers contrastive clues from two fine-grained images and attentively distinguish them via pair interaction while another approach of branching into multiple focus of an image in the Multi-branch and Multi-scale Attention Learning for Fine-Grained Visual Categorization (MMAL-Net) [24]

This paper will focus on fine-grained visual categorization (FGVC). FGVC, on the other hand, differs from standard object detection in that there is more intra-class variance than inter-class variance. This suggests that there is not much of a distinction between the categories, and they appear to be nearly identical. As a result, discovering discriminative features within similar classes is difficult in this area of research.

Aircrafts are considered one of the FGVC's objects along with other FGVC such as birds, pets, flowers, and cars. With aircraft designs improving across decades, with various models and variants introduced and delivered from each aircraft manufacturer, aircraft model recognition has been increasingly challenging and requires more effort to achieve better accuracy. The structure of aircraft changes for different architecture such as wing shapes, number of wings, engines, number of wheels, etc. This particular variation does not share the same with categories such as animals [21]. With custom design and make for different airlines, a similar aircraft model for different airlines will have different look(livery) or a different look for a similar aircraft model will coexist in an airline.

In this paper, Section II explores some fine-grained visual classification related works. The proposed work is explained in Section III. Then, the experiments and results are explored in Section IV. Finally, Section V concludes the paper.

II. Related Work

There are a lot of research works have been proposed for fine-grained visual classification. In this section, three recent works that achieve superior or state-of-the-art performance in the FGVC will be summarized and briefly mentioned some methods.

The spatial transformer network [25] proposed by Jaderberg et al. allows the neural network to be able to actively spatially transform feature maps with no additional modification to the optimization process. While multi-attention convolutional neural network (MA-CNN) [20] generates parts from features learning and classifies the image by each part by their probability scores. Yang et al. [26] utilize the proposed self-supervision mechanism to localize the features on the image itself. The self-supervision is mainly driven by the Navigator-Teacher-Scrutinizer agents which every agent supports each other, was called NTS-Net. Zheng et al. then proposed the trilinear attention sampling network(TASN) [27] uses attention maps generated from feature maps and extract details from it for further optimization. One recent work generates gates by comparing mutual vector, which is learned from the semantic difference between two input images, with individual vectors. This gives the Attentive Pairwise Interaction Network(API-Net) [23] the ability to capture information by pairwise interaction between two images. Another recent work [28] introduce an efficient end-to-end localization and achieved a high recognition accuracy in the FGVC and achieved the second highest accuracy in FGVC Aircraft after TBMSL-Net [24]. TBMSL-Net or three-branch and multi-scale learning use a raw image, object image generated through attention object location module(AOLM), and parts image generated through attention part proposal module(APPM) to train in three different "branches".

A. Attentive Pairwise Interaction Network (API-Net) [23]

This proposed API-Net by Zhuang, Wang, and Qiao [23] inspired by human behavior instead of learning from one image, we often compare image pairs of the same fine-grained objects and learn the subtle differences between them. Thus, API-Net, like humans, can constructively recognize two fine-grained images by looking for constructive clues from them and attentively distinguishing them by pair interaction [23]. These consist of three submodules in API-Net which play important roles in progressively distinguishing pair image input, they are mutual vector learning, gate vector generation, and the pairwise interaction.

In the mutual vector learning module, a mutual vector Xm is learned using a multilayer perceptron(MLP) by extracting the feature vector from two input images X1 and X2. In gate vector generation, further comparison between Xm with X1 and X2 to discover gi, which stores the distinctive of each image, as a result of performing Hadamard product between each image with Xm and gone through a sigmoid function. Zhang et al. have the pairwise interaction inspired by human behavior not only checking the image's prominent parts but also with distinct parts from the other image. This results in two feature vectors being produced via residual attention for each image from its gate vector and also from its pair's gate vector to make the clues more obvious for each image. The authors then introduce a SoftMax classifier to predict the image into each respective category with training loss to learn more for API.

B. Efficient End-to-end Localization [28]

This proposed method by Hanselmann and Ney [28] eliminates the approach used by other localization methods which

are inefficient in the end-to-end setup. This proposed network utilized the AttNet, AffNet, and classification network in an end-to-end localization. Attention map derived from the input image processed in AttNet will be passed to AffNet to define the bounding box. Then, bilinear sampling is applied to do the cropping and then the cropped image is sent to the classification network. Figure 1 shows the The architecture of the End-to-End Localization model.

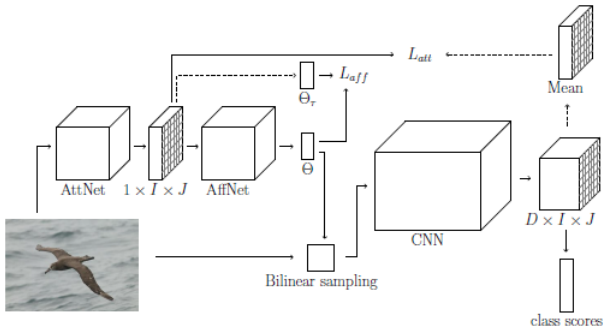

Fig. 1. The architecture of the End-to-End Localization model. Dashed arrows show supervision signals for self-supervised losses thus no gradient flows back [28]

AttNet in this End-to-end Localization model is lightweight and efficient as stated by the author, and it will predict an attention map for AffNet. AttNet will learn from its local loss.

C. Three-Branch and Mutil-Scale Learning (TBMSL-Net) [24]

The TBMSL-Net is designed like its given name, which consists of three branches in the network design. Each branch is like a standalone small network and can work on its own but at the same time work closely with each other. The parameters of the convolutional network and the fully connected network in these three branches are shared and this helps increase the classification accuracy and at the same time reduce calculation. TBMSL-Net is not only able to recognize the discriminative region but it can also be trained end-to-end. There consists of two modules in TBMSL-Net, playing a very important role in the network and at the same time connecting the branches. The first module is the Attention Object Location Module(AOLM) and the second module is the Attention Part Proposal Module(APPM). From the feature maps output from the convolutional network of the input raw image, AOLM will acquire the object's location information and crop them from the raw image. Then, the feature maps output from the object image is gone through APPM to locate the information of parts that consist of distinct features. Then the object image is sent down to have it crop into part images. Figure 2 shows the framework of the TBMSL-Net.

All three branches use cross-entropy loss as the classification loss. The total loss, which is the addition of all three losses, works together to allow classification predictions make based on object characteristics or part's fine-grained characteristics. This optimizes the performance of the TBMSL-Net.

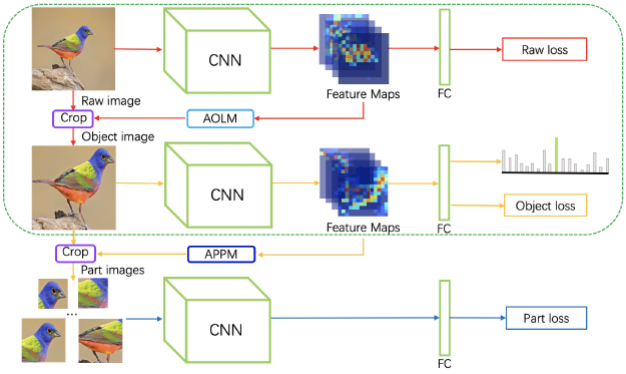

Fig. 2. The framework of the TBMSL-Net [24]

So far, TBMSL-Net achieved the best performance in FGVC-Aircraft which is 94.5/

Beside the above explained related works, Table 1 summarize the results of some of the state-of-the art works on on FGVC-Aircraft.

III. PROPOSED DESIGN

A. Database

Aircraft, like birds and automobiles, are fine-grained objects. Identifying aeroplanes is difficult for a computer due to the numerous models and designs created over decades. A benchmark dataset for FGVC is Fine-Grained Visual Classification of Aircraft (FGVC-Aircraft) [21]. This information was used in the FGComp 2013 fine-grained recognition challenge, which ran concurrently with the ImageNet Challenge 2013. This collection contains 10,200 aeroplane photographs, including 100 images for each of the 102 aircraft model variants. About 60–70% of the total photographs will be used for training, with the remaining images being used for testing.

B. Implementation

The Attentive Pairwise Interaction Network (API-Net) employed in this study for fine-grained aircraft image classification is discussed in this section. The structure of the API-Net is described in depth first, followed by the modifications made to the entire learning process using the API-Net. Instead of learning the visual importance from a single image, the Attentive Pairwise Interactive Network evaluates two images at the same time and discovers the significant distinctions between them, as mentioned in the literature. This design was inspired by how humans frequently distinguish similar objects by comparing image pairs to determine the small differences. The original work used three different convolutional neural network backbones in the network and released the API-Net with the highest accuracy from those three separate CNN implementations.

ResNet-50, ResNet-101, and DenseNet-161 are the backbones recommended in the literature for installation. The degradation was addressed by ResNet by incorporating a deep residual learning architecture into their research. ResNet's core

TABLE I
LIST OF RECENT STATE-OF-ART PERFORMANCE WORK ON FGVC-AIRCRAFT

Paper	Model	Year	Accuracy(%)
Three-branch and Mutil-scale learning for Fine-grained Image Recognition (TBMSL-Net) [24]	TBMSL-Net	2020	94.5
Fine-Grained Visual Classification with Efficient End-to-end Localization [28]	AttNet & AffNet	2020	94.1
Learning Attentive Pairwise Interaction for Fine-Grained Classification [23]	API-Net	2020	93.9
Weakly Supervised Fine-Grained Image Classification via Gaussian Mixture Model-Oriented Discriminative Learning [29]	DF-GMM	2020	93.8
Fine-grained Recognition: Accounting for Subtle Differences between Similar Classes [22]	DB	2019	93.5
Fine-Grained Visual Classification with Batch Confusion Norm [30]	BCN	2019	93.5
ELoPE: Fine-Grained Visual Classification with Efficient Localization, Pooling and Embedding [31]	ELoPE	2019	93.5
Fine-Grained Visual Classification via Progressive Multi-Granularity Training of Jigsaw Patches [32]	PMG	2020	93.4
Channel Interaction Networks for Fine-Grained Image Categorization [33]	CIN	2020	93.3
Graph-propagation based Correlation Learning for Weakly Supervised Fine-grained Image Classification [34]	GCL	2020	93.2

concept is an identity shortcut connection, which skips one or more levels and performs identity mapping after each block, with the identity output layered with the output of the stacked layers. As depicted in Figure 3, the identification shortcut connection building blocks.

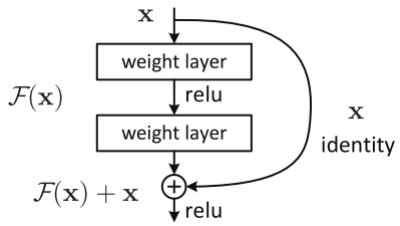

Fig. 3. A residual building block

While on the other hand, Huang et al. proposed DenseNet further exploits the shortcut connections and utilizing a different connectivity pattern. In DenseNet, it connects all layers directly with each other comparing to ResNet which is done only on a building block. Consequently, a layer will aggregate the feature maps of all preceding layers as input.

IV. EXPERIMENTS AND RESULTS

A. Experiments without data augmentations

Figure 4 depicts the outcomes of research reported in the literature. Using DenseNet-161 as the backbone CNN with 30 class sizes and 4 image size settings, the literature achieved the greatest accuracy of 93.9 percent in FGVC-Aircraft. While the second top result, 93.4%, was reached by using ResNet-101 as the backbone CNN with the same 30 class size and 4 image size parameter. The literature achieved a little lower number, 93.0 percent, when ResNet-50 was used as the backbone CNN with 30 class size and 4 image size configuration. When comparing the literature's settings to mine, the class/image size setting remains 30/4, despite the literature's use of ResNet-50.

We can reach 92.77% accuracy by using the literature's original ResNet-50 setup and only tweaking the class/image size to 10 class size and 2 image size. This score will be used as the baseline for the next experiment in this study.

Method	Backbone	Extra S.	Aircraft
BoT (Wang et al. 2016)	VGGNet-16	Yes	88.4
MG-CNN (Wang et al. 2015)	VGGNet-19	Yes	86.6
KP (Cui et al. 2017)	VGGNet-16	No	86.9
LRBP (Kong and Fowlkes 2017)		No	87.3
G^2DeNet (Wang, Li, and Zhang 2017)		No	89.0
Grassmann Pool(Wei et al. 2018)		No	89.8
HBP (Yu et al. 2018)		No	90.3
DFL-CNN (Wang, Morariu, and Davis 2018)		No	92.0
B-CNN (Lin, RoyChowdhury, and Maji 2015)	VGGNet-19	No	84.1
RACNN (Fu, Zheng, and Mei 2017)		No	88.4
MACNN (Zheng et al. 2017)		No	89.9
Deep KSPD (Engin et al. 2018)		No	91.5
NTS-Net (Yang et al. 2018)	ResNet-50	No	91.4
iSQRT-COV (Li et al. 2018b)	ResNet-101	No	91.4
PC (Dubey et al. 2018a)	DenseNet-161	No	89.2
MaxEnt (Dubey et al. 2018b)		No	89.8
Our API-Net	ResNet-50	No	93.0
Our API-Net	ResNet-101	No	93.4
Our API-Net	DenseNet-161	No	**93.9**

Fig. 4. Comparison with The-State-of-The-Art on the FGVC Aircraft. Where Extra S. stands for Extra Supervision

B. Experiment with data augmentations

We have obtained the benchmark score after obtaining the API-Net score using ResNet-50 with 10/2 class/image size. We will compare the results of each experiment with data augmentation to the benchmark score in the Figure 4. It is worth noting that the class size and image size will be set to 10 and 2 in all subsequent trials. Begin by adjusting the brightness and contrast to 0.2. This setting yielded a 92.74% result, which is somewhat lower than the benchmark score. After that, we added grayscale to the data augmentation in the second experimental setting, which kept the brightness and contrast values the same. When the grayscale value is set to 0.3, the system will select 30 percent of the whole training image at random and convert it to grayscale for an epoch.

Fortunately, we can see that this setting yields a somewhat higher result of 92.86%. In the second experiment, we kept everything the same but changed the grayscale from 0.3 to 0.5 and got a reading of 92.8%. Things improved when we deleted the brightness and contrast from the data augmentation, leaving only the grayscale. For one new trial, we used grayscale 0.3, while for another experiment, we used grayscale

0.5. For the first time, the resulting percentage exceeded 93%, with 93.25% for grayscale 0.3 and 93.46% for grayscale 0.5.

Various studies have shown that data augmentation can assist improve network accuracy, particularly on data-hungry networks. As a result, in this study, we will integrate data augmentation into API-Net to see if data augmentation applied to the API-Net improves the network's accuracy. The following are examples of data augmentation used in this paper:

- Colour Jitter
- Grayscale

With these two data augmentation, we have empirically tried with different setting on the API-Net. These settings are:

- The first setting: Colour Jitter: Brightness = 0.2, Contrast = 0.2, Grayscale = 0.3.
- The second setting: Colour Jitter: Brightness = 0.2, Contrast = 0.2, Grayscale = 0.3.
- The third setting: Colour Jitter: Brightness = 0.2, contrast = 0.2.
- The fourth setting: Grayscale = 0.3.
- The fifth setting: Grayscale = 0.5

With a variety of settings planned for the API-Net, we had the ability to see which ones helped boost accuracy the most. Furthermore, in the MMAL-Netz [24], the colour jitter with brightness = 0.2 and contrast = 0.2 has been shown to improve accuracy. We utilised the grayscale settings of 0.3 and 0.5 to see which one might obtain the highest level of accuracy. Figures 5 and 6 show the API-Net networks with colour jitter and grayscale.

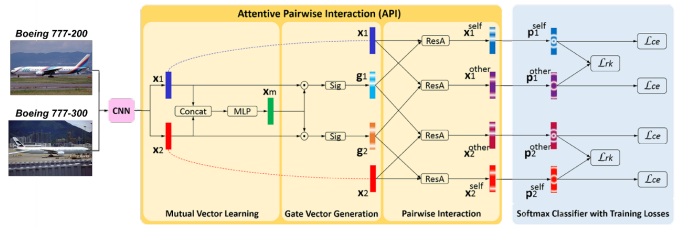

Fig. 5. API-Net with introduced colour jitter

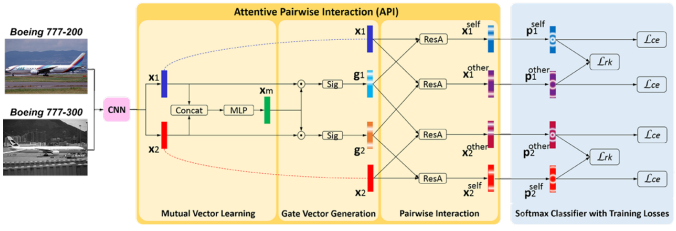

Fig. 6. API-Net with introduced grayscale

Every image is first scaled to 512×512 pixels, then cropped with random cropping in training and centre cropping in testing to 448×448 pixels. After that, the image is randomly modified in colour jitter, and the random image is turned into grayscale. ResNet-50 was used as the CNN backbone, and a random sample of 10 categories was taken in a batch, with two images randomly selected from each category. The literature's architecture, including the completely connected layer, is fully employed in the network architecture. Learning rate, momentum, and weight decay are all fully transferred from the literature. With a learning rate of 0.01, SGD has 0.9 momentum and 0.0005 weight decay. Using cosine annealing as the function, the learning rate will be changed. We will train for 100 epochs in each setting.

Table II shows the results of different setting in comparison to the literature's result under the similar CNN backbone.

TABLE II
COMPARISON OF DIFFERENT SETTING WITH THE LITERATURE'S ORIGINAL VERSION

Backbone	Class/Image size	Data Augmentation	Accuracy
ResNet-50	30/4*	Nil	93.0*
ResNet-50	10/2**	Nil	92.77**
ResNet-50	**10/2**	**Brightness = 0.2 Contrast = 0.2 Grayscale = 0.3**	**92.86**
ResNet-50	**10/2**	**Brightness = 0.2 Contrast = 0.2 Grayscale = 0.5**	**92.8**
ResNet-50	**10/2**	**Brightness = 0.2 Contrast = 0.2**	**92.74**
ResNet-50	**10/2**	**Grayscale = 0.3**	**93.25**
ResNet-50	**10/2**	**Grayscale = 0.5**	**93.46**
ResNet-101*	30/4*	Nil	93.4*
DenseNet-161*	30/4*	Nil	93.9*

*The literature's setting on class/image size and the respective accuracy
**The literature's setting with class/image size changed to 10/2 and the respective accuracy. This result will be the benchmark of the following work in the study

We can see that data augmentation increases overall accuracy marginally, as indicated in Table II. In the table, we can see a special value of 92.74%, which is a tiny decrease from the previous API-Net option of 10/2 class/image size, which was 92.77%. Furthermore, we see that data augmentation with a brightness and grayscale pair does not result in an accuracy increase of more than 93.0%, although data augmentation with grayscale alone is capable of exceeding 93.0% accuracy. It is a little unusual that adjusting the brightness and contrast of the images won't help the API-Net, but introducing grayscale will. With this result, we can demonstrate that brightness and contrast do not aid API-Net training. While, if we introduce a longer training time, which will be our future work, the accuracy may be higher than the one shown in the table.

V. CONCLUSION

In this study, data augmentation was added to an existing CNN for FGVC-Aircraft. API-Net is a pre-existing CNN that was used in this investigation. The study concluded that data augmentation does help to improve the API-Net when tweaking with various data augmentation settings. Brightness and

contrast adjustment, on the other hand, are ineffective in the research context, and they have reduced the network's potential to achieve higher accuracy. Grayscale, on the other hand, aids in boosting the API-accuracy. Net's API-Net requires a huge load of GPU to train using the settings in the literature, hence this work is primarily constrained by the hardware setup. Although we can demonstrate the improvement of the API-Net in this study, we can push the API-accuracy Net's higher than the literature's if given a more powerful machine.

ACKNOWLEDGMENT

This paper is partially supported by Universiti Malaysia Pahang research Grant (RDU192212).

REFERENCES

[1] N. Islam, Z. Islam, and N. Noor, "A survey on optical character recognition system," *arXiv preprint arXiv:1710.05703*, 2017.

[2] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, no. 2, pp. 261–318, 2020.

[3] K. Sundararajan and D. L. Woodard, "Deep learning for biometrics: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 3, pp. 1–34, 2018.

[4] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[7] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[9] J. Greenhalgh and M. Mirmehdi, "Real-time detection and recognition of road traffic signs," *IEEE transactions on intelligent transportation systems*, vol. 13, no. 4, pp. 1498–1506, 2012.

[10] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 12, pp. 2179–2195, 2008.

[11] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2011.

[12] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection: A review," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 5, pp. 694–711, 2006.

[13] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: past, present and future," *Computer Vision and Image Understanding*, vol. 138, pp. 1–24, 2015.

[14] R. Bagasjvara, I. Candradewi, S. Hartati, and A. Harjoko, "Automated detection and classification techniques of acute leukemia using image processing: A review," in *2016 2nd International Conference on Science and Technology-Computer (ICST)*. IEEE, 2016, pp. 35–43.

[15] E. Jana, R. Subban, and S. Saraswathi, "Research on skin cancer cell detection using image processing," in *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE, 2017, pp. 1–8.

[16] A. Dutta and A. Dubey, "Detection of liver cancer using image processing techniques," in *2019 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2019, pp. 0315–0318.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[19] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4438–4446.

[20] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5209–5217.

[21] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.

[22] G. Sun, H. Cholakkal, S. Khan, F. Khan, and L. Shao, "Fine-grained recognition: Accounting for subtle differences between similar classes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 047–12 054.

[23] P. Zhuang, Y. Wang, and Y. Qiao, "Learning attentive pairwise interaction for fine-grained classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 130–13 137.

[24] F. Zhang, G. Zhai, M. Li, and Y. Liu, "Three-branch and mutil-scale learning for fine-grained image recognition (tbmsl-net)," *arXiv preprint arXiv:2003.09150*, 2020.

[25] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, pp. 2017–2025, 2015.

[26] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 420–435.

[27] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5012–5021.

[28] H. Hanselmann and H. Ney, "Fine-grained visual classification with efficient end-to-end localization," *arXiv preprint arXiv:2005.05123*, 2020.

[29] Z. Wang, S. Wang, P. Zhang, H. Li, W. Zhong, and J. Li, "Weakly supervised fine-grained image classification via correlation-guided discriminative learning," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1851–1860.

[30] Y.-C. Hsu, C.-Y. Hong, D.-J. Chen, M.-S. Lee, D. Geiger, and T.-L. Liu, "Fine-grained visual recognition with batch confusion norm," *arXiv e-prints*, pp. arXiv–1910, 2019.

[31] H. Hanselmann and H. Ney, "Elope: Fine-grained visual classification with efficient localization, pooling and embedding," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1247–1256.

[32] R. Du, D. Chang, A. K. Bhunia, J. Xie, Z. Ma, Y.-Z. Song, and J. Guo, "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *European Conference on Computer Vision*. Springer, 2020, pp. 153–168.

[33] Y. Gao, X. Han, X. Wang, W. Huang, and M. Scott, "Channel interaction networks for fine-grained image categorization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 818–10 825.

[34] Z. Wang, S. Wang, H. Li, Z. Dou, and J. Li, "Graph-propagation based correlation learning for weakly supervised fine-grained image classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 289–12 296.