Scientific Research Publishing

Journal of Computer and Communications, 2024, 12, 307-316
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Progressive Feature Fusion-Based Manhole Cover Defect Recognition Method

Tingting Hu[1], Xiangyu Ren[1], Wanfa Sun[1], Shengying Yang[1*], Boyang Feng[2]

[1]School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, China
[2]Department of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia
Email: *syyang@zust.edu.cn

How to cite this paper: Hu, T.T., Ren, X.Y., Sun, W.F., Yang, S.Y. and Feng, B.Y. (2024) A Progressive Feature Fusion-Based Manhole Cover Defect Recognition Method. *Journal of Computer and Communications*, **12**, 307-316.
https://doi.org/10.4236/jcc.2024.128019

Received: July 29, 2024
Accepted: August 27, 2024
Published: August 30, 2024

Abstract

Manhole cover defect recognition is of significant practical importance as it can accurately identify damaged or missing covers, enabling timely replacement and maintenance. Traditional manhole cover detection techniques primarily focus on detecting the presence of covers rather than classifying the types of defects. However, manhole cover defects exhibit small inter-class feature differences and large intra-class feature variations, which makes their recognition challenging. To improve the classification of manhole cover defect types, we propose a Progressive Dual-Branch Feature Fusion Network (PDBFFN). The baseline backbone network adopts a multi-stage hierarchical architecture design using ResNet50 as the visual feature extractor, from which both local and global information is obtained. Additionally, a Feature Enhancement Module (FEM) and a Fusion Module (FM) are introduced to enhance the network's ability to learn critical features. Experimental results demonstrate that our model achieves a classification accuracy of 82.6% on a manhole cover defect dataset, outperforming several state-of-the-art fine-grained image classification models.

Keywords

Feature Enhancement, Progressive, Dual-Branch, Feature Fusion

1. Introduction

Manhole covers, located outdoors, are exposed to wind and rain and operate in complex road environments, which can lead to issues such as damage, protrusion, absence, and problems with the well ring. Moreover, some unscrupulous manufacturers produce manhole covers of inconsistent quality in an effort to reduce costs, creating numerous safety hazards. Existing studies have explored the

installation of sensors within manhole covers to achieve smart manhole cover systems, but the high construction and maintenance costs associated with these sensors prevent their widespread deployment.

In recent years, manhole cover recognition technologies based on computer vision have received extensive attention and application. Unlike manual identification methods, deep learning approaches are less influenced by human subjectivity and offer more accurate and faster recognition, enabling early detection of manhole cover damage and absence. Currently, traditional image detection methods mainly focus on detecting the presence of manhole covers. For example, Wei *et al.* [1] proposed a method based on Support Vector Machines (SVM) combined with oriented gradient histograms that possess symmetry features, using laser point cloud intensity images and ground orthophotos to accurately detect manhole covers. Qiao *et al.* [2] utilized a multi-feature fusion approach, leveraging prior knowledge to define the detection range and combining gray-level features, edge features, and symmetry features to determine the exact location of manhole covers. Ji *et al.* [3] presented a street manhole cover detection method combining multi-view matching with feature extraction. They performed scene segmentation using LiDAR data, extracting regions of interest containing arcs, and further checked whether they were manhole covers. Image detection technologies analyze and process data more objectively and are less influenced by subjective factors; however, they perform poorly in recognizing the types of manhole cover defects. Moreover, when the texture of manhole cover defects is similar to the road background, surface features are difficult to distinguish from the surrounding pavement, leading to missed detections or false positives during detection. Therefore, to address these issues, we propose a Progressive Dual-Branch Feature Fusion Network (PDBFFN) based on the ResNet50 network to improve the model's recognition of manhole cover defect types. The main contributions of this paper are as follows:

- We propose an end-to-end framework that integrates attention mechanisms and fine-grained feature learning. This framework collaboratively integrates these features for fine-grained feature learning, enabling better recognition of manhole cover defect classifications.
- We propose a novel Feature Enhancement Module (FEM) that captures the most discriminative features at each step and obtains image information at different levels, enabling multi-scale progressive feature fusion and yielding richer category information for manhole cover defects.
- We crawled relevant manhole cover images from the web and expanded the dataset through data augmentation techniques, constructing a manhole cover defect dataset consisting of 5 defect categories with a total of 9654 images. The proposed progressive fusion network achieved an accuracy of 82.6% on this dataset.

2. Related Works

2.1. Fine-Grained Image Classification

Currently, mainstream fine-grained classification methods include localization-

based methods [4] [5] and feature encoding methods [6]. The former utilizes attention mechanisms, clustering, and other techniques to discover distinctive regions, while the latter captures more subtle regional features by computing higher-order information. Strongly supervised localization methods require expensive and time-consuming manual annotations, whereas weakly supervised methods can be trained using only classification labels, thus receiving more attention. The advent of attention mechanisms has further enhanced the performance of image classification models, with mainstream classification models increasingly focusing on diverse attention-guided methods. CAL [4] uses counterfactual intervention to encourage the network to learn more attention regions. Contrastive input batch construction methods [7] strengthen the ability of features to contain discriminative information by calculating cues between different features. Feature enhancement and suppression methods [8] force the network to uncover latent information by enhancing or suppressing key parts and local regions.

2.2. Feature Fusion

Learning features from local regions plays a crucial role in fine-grained image classification tasks. Utilizing attention mechanisms to extract local features has become a trend. However, these methods have two main limitations: first, they often focus on the most prominent parts of the image, neglecting other areas. Second, they treat different local features separately, ignoring the connections between them. Networks with a multi-stage hierarchical architecture design can typically capture features at different scales, where these feature maps contain different information emphases. Lower-level features can capture more detailed information and focus on more critical regions, such as edge textures, shapes, and colors, while higher-level features contain richer semantic information and focus on the target region as a whole. Therefore, effectively fusing features at different scales can enhance the model's feature representation capabilities and improve its recognition performance. FPN [9] achieved significant success in object detection by aggregating feature maps from different layers. However, using element-wise addition as the aggregation operation still has limited functionality. Wang *et al.* [10] proposed a non-local operation, where the response at a spatial position is computed as the weighted sum of all positions in the feature map. SG-Net [11] used non-local operations to fuse feature maps from different layers. Addressing the aforementioned issues, in this paper, we build upon the ResNet50 backbone network. First, we enhance the information representation at each stage through a Feature Enhancement Module (FEM). Then, we obtain image information at different levels of the model and perform progressive dual-branch feature fusion, allowing the model to focus on global information while extracting, amplifying, and associating local information. This promotes the model to learn richer feature representations and enhances its discriminative power.

3. Methods

3.1. Overall Network Structure

The overall network structure is shown in **Figure 1**. In this paper, we use the ResNet50 backbone network as the feature extractor for fine-grained images, cascading the extraction of visual features from shallow to deep stages. At the end of the second, third, and fourth stages, the output feature maps are fed into the Feature Boosting Module (FBM) for feature enhancement operations. Then, the Gradual Feature Fusion Module (GFFM) is used to perform multi-scale feature fusion operations on the feature maps from different stages, enabling the network to learn richer image feature information.

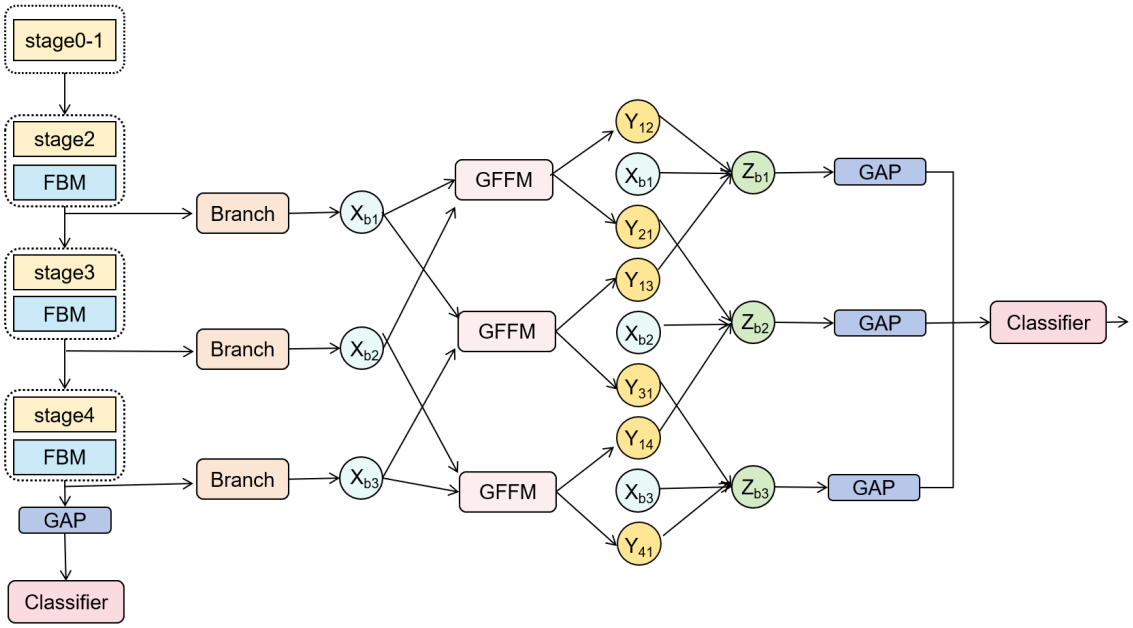

Figure 1. Overall framework diagram.

3.2. Feature Enhancement Module

First, given a feature map $X \in R^{C\times W\times H}$ from a specific layer, where C, W, H represent the number of channels, width, and height, respectively. We simply split X uniformly along the width dimension into k parts, and denote each stripe part as $X_i \in R^{C\times(W/k)\times H}$, where $i \in [1,k]$. Then, we use a 1 × 1 convolution φ to explore the importance of each part:

$$A_i = Relu\left(\varphi\left(X_i\right)\right) \in R_1 \times \frac{W}{k} \times H \tag{1}$$

The non-linear function ReLU is used to remove negative activations. φ is shared among the different stripe parts and acts as a splitter. Then, the average of A_i is taken as the importance factor b_i' of X_i.

$$b_i' = GAP\left(A_i\right) \tag{2}$$

where GAP stands for Global Average Pooling. The softmax function is used to

normalize $B' = \left(b_1', b_2', \cdots, b_k'\right)^T$.

$$b_i = \frac{\exp\left(b_i'\right)}{\sum_{j\in[1,k]}\exp\left(b_j'\right)} \tag{3}$$

Through the normalized importance factor $B' = \left(b_1', b_2', \cdots, b_k'\right)^T$, the most significant part can be immediately determined. Then, by enhancing the most significant part, the boosted feature X_b is obtained:

$$X_b = X + \alpha * \left(B \otimes X\right) \tag{4}$$

where α is a hyperparameter that controls the degree of enhancement, and $\otimes$ denotes element-wise multiplication. The structure of the feature enhancement is shown in **Figure 2**.

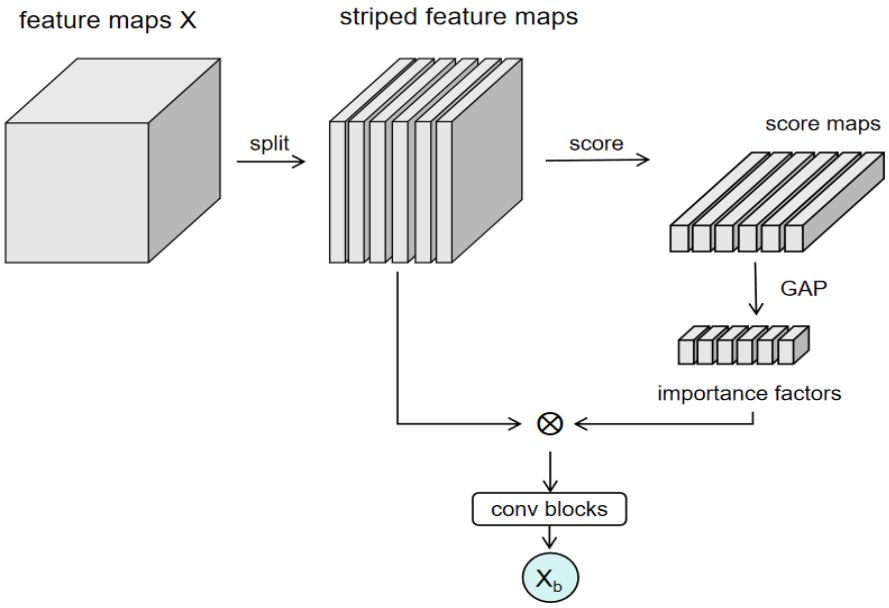

Figure 2. Feature enhancement module.

3.3. Gradual Feature Fusion Module

After extracting three levels of feature maps from the enhanced network, each feature map is fed into a corresponding processing branch. The processing branch consists of a series of convolutional layers. One column of the output is multi-channel, while the other is an attention branch that outputs a single channel. The two columns of feature maps are multiplied together to form the output of the branch.

Since each stage's feature map contains different feature information, in order to more comprehensively mine the image feature information, feature maps from any two stages $X_{bi} \in R^{H_i \times W_i \times C}$ and $X_{bj} \in R^{H_j \times W_j \times C}$ are selected and fed into the feature fusion module. The two feature matrices are multiplied to obtain a similarity matrix $M \in R^{H_i W_i \times H_j W_j}$.

$$M = F\left(X_{bi}, X_{bj}\right), F\left(X, Y\right) = X^T Y \tag{5}$$

where the lower the similarity, the more complementary information is indicated. The complementarity correlation matrix ($-M$) is obtained by taking the inverse of the similarity matrix.

Subsequently, the complementarity correlation matrix is normalized using the Softmax operation. It is then multiplied with the feature maps of the two stages to obtain complementary output feature maps Y_{ij} and Y_{ji}. The calculation process is formulated as follows:

$$A_{ij} = softmax\left(-M^T\right) \in [0,1]. \tag{6}$$

$$A_{ji} = softmax\left(-M\right) \in [0,1] \tag{7}$$

$$Y_{ij} = X_{bi} A_{ij} \in R^{H_j \times W_j \times C} \tag{8}$$

$$Y_{ji} = X_{bj} A_{ji} \in R^{H_i \times W_i \times C} \tag{9}$$

In the equation, $A_{ij} \in R^{H_j W_j \times H_i W_i}$ and $A_{ji} \in R^{H_i W_i \times H_j W_j}$ represent the indices of the feature maps. For Y_{ij}, it is the complementary information output feature map of X_{bi} relative to X_{bj} and similarly for Y_{ji}. The structure diagram of the feature fusion module is shown in **Figure 3**.

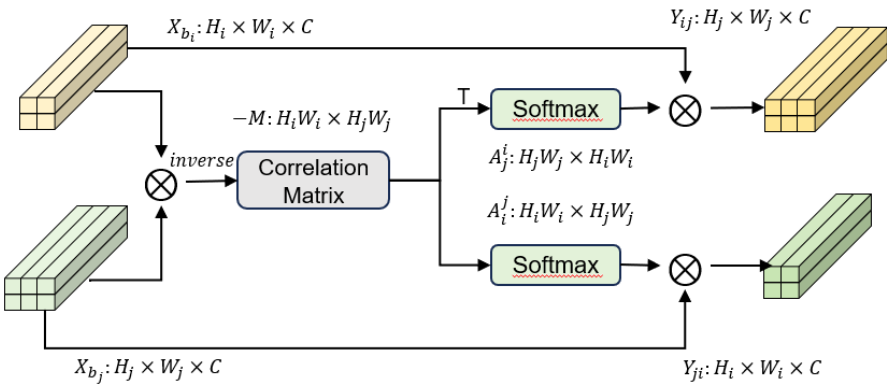

Figure 3. Feature fusion module.

To make the extracted feature representations more diverse, the input feature maps from each stage are fused with the output complementary feature maps, resulting in the final output feature map Z_{bi}.

$$Z_{bi} = X_{bi} + \sum_{j=1}^{3} Y_{ji} \left(j \neq i\right) \tag{10}$$

4. Experiments

4.1. Implementation Details

The system experimental environment is Ubuntu 18.04, with a software configuration including CUDA 12.2 and Python 3.9. The hardware configuration includes an NVIDIA RTX 3080 GPU with 12 GB of VRAM. The model training platform is the PyTorch deep learning framework based on the Python programming language. In the experiments, all training images are first resized to 256 × 256 and then randomly cropped to 224 × 224. Data augmentation techniques,

including random rotation, random horizontal flipping, and random image enhancement, are applied to expand the training dataset. For the test set images, only resizing to 224 × 224 is performed. Finally, a unified mean normalization is applied to both the dataset and the training set to enhance the data representation. The batch size is set to 8. The official ResNet50 model is used as the backbone network, initialized with pre-trained weights, while the weights of the added modules are randomly initialized. All models are trained using the Adam optimizer with a weight decay of 5e−4. The learning rate for the backbone network is 1e−4, and the learning rate for the added networks is 1e−3, with dynamic adjustment of the learning rate through cosine annealing (CosineAnnealing).

Accuracy is selected as the evaluation metric for the model, defined as:

$$\mathrm{Acc} = \frac{\mathrm{I}_{\mathrm{ac}}}{\mathrm{I}_{\mathrm{total}}} \tag{11}$$

where I_{ac} is the number of correctly classified images, and $\mathrm{I}_{\mathrm{total}}$ is the total number of images in the test set.

4.2. Datasets and Metrics

The dataset used in this paper is the manhole cover defect dataset, for which we crawled over 2000 manhole cover photos from the internet. The manhole covers are categorized into 5 classes, as shown in **Table 1**. To improve the model's generalization capability and prevent overfitting, this experiment uses data augmentation to expand the dataset to 9654 images, including 7256 images for the training set and 2398 images for the test set.

Table 1. Types of manhole cover defects.

Sort	Number
Good	2519
Broke	1613
Lose	2336
Uncover	2455
Circle	732

4.3. Ablation Studies

To validate the effectiveness of each module in the network, we conducted ablation experiments on the dataset. We sequentially introduced the Feature Boosting Module (FBM) and the Gradual Feature Fusion Module (GFFM) on top of the backbone network and trained them individually. The experimental results are shown in **Table 2**. The ResNet50 backbone network can achieve a classification accuracy of 75.85%. With the introduction of the Feature Boosting Module, the model's performance improves by 3.09 percentage points, achieving a classification accuracy of 78.94%. Further introducing the Gradual Feature Fusion Module

increases the model's performance by another 3.6 percentage points, reaching a classification accuracy of 82.6%. The experimental results indicate that each module effectively enhances the overall classification performance of the model.

Table 2. Analysis of ablation experiment.

NO	Method	Model Composition	Accuracy/%
1	Resnet50 (baseline)	Resnet	75.85
2	ResFM	Resnet + FBM	78.94
3	ResFM	Resnet + FBM + GFFM	82.6

4.4. Quantitative Comparison

To further validate the superiority of our method, we selected popular image classification methods such as ViT and Swin and conducted comparative experiments on the manhole cover defect dataset. The results are shown in **Table 3**. Our model outperforms the DenseNet model by 28.5%, the MobileNet model by 20%, and the Res2Net model by 17.1%. Compared to Swin-Transformer and ViT, our model also demonstrates superior classification performance. Overall, our model shows higher fine-grained image classification accuracy compared to Transformer-based models, especially in scenarios with limited data.

Table 3. Accuracy on different weakly supervised fine-grained image classification models.

Model	Accuracy/%
DenseNet	54.083
MobileNet	62.583
Res2Net	65.458
Swin-transformer	43.375
ViT	40.292
NTS-Net	63.4
Ours	82.6

4.5. Visualization

To gain a more intuitive understanding of the role of each module and to further validate the credibility of our method, we conducted visualization experiments. We analyzed the feature maps of the backbone network ResNet50, as well as those after introducing the FBM and GFFM modules. The results are shown in **Figure 4**.

It can be observed that for the backbone network ResNet50, the model only focuses on some prominent features and suffers from misalignment issues, such as in the case of the 'Uncover' manhole cover, where the backbone network does not pay attention to the relevant features of the manhole cover defect. After

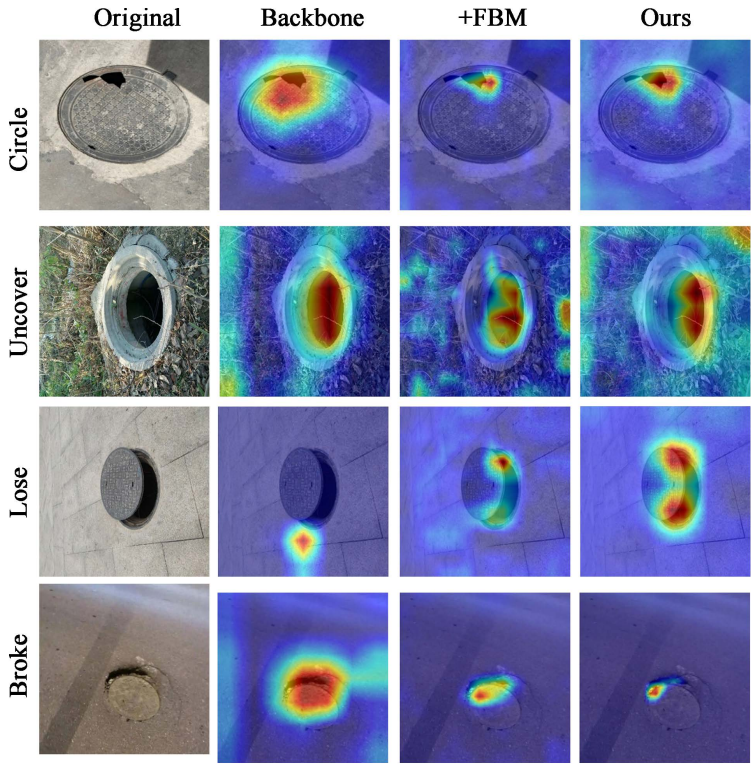

Figure 4. Visualization of heat maps of different categories.

introducing the FBM module, the model reduces its focus on irrelevant regions and is able to locate the regions of interest. After introducing the GFFM module, the model learns richer image features and pays more attention to the regions of interest.

5. Conclusion

To address the scarcity of research on category division for manhole cover defects in computer vision networks and the poor performance of classification recognition, we propose a fine-grained image classification model that incorporates feature enhancement and gradual dual-branch feature fusion. By adopting a feature enhancement module, the network is encouraged to learn more salient feature information, improving the model's ability to extract detail features while reducing interference from image background noise. Through the introduction of gradual feature fusion, the network learns more comprehensive and diverse feature information. Experimental results show that the proposed model outperforms most mainstream methods. In future work, we will continue to investigate the use of visual Transformer architectures in fine-grained image classification, aiming to optimize network architecture and simplify layer numbers without sacrificing performance, exploring networks that are more suitable for fine-grained image classification.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Wei, Z., Yang, M., Wang, L., *et al.* (2019) Customized Mobile LiDAR System for Manhole Cover Detection and Identification. *Sensors*, **19**, Article 2422. https://doi.org/10.3390/s19102422

[2] Qiang, R.P., Sun, H. and Dong, Y.C. (2018) Implementation of Manhole Cover Detection System Based on Multi Feature Fusion. *Electron Technology Applications*, **44**, 44-47.

[3] Oji, S., Shi, Y. and Shi, Z. (2012) Manhole Cover Detection Using Vehicle-Based Multi-Sensor Data. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, **XXXIX-B3**, 281-284. https://doi.org/10.1109/9.402235

[4] Rao, Y., Chen, G., Lu, J., *et al.* (2021) Counterfactual Attention Learning for fine-Grained Visual Categorization and Reidentification. *Proceedings of the* 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 1005-1014. https://doi.org/10.1109/9.402235

[5] Liu, C., Xie, H., Zha, Z., Ma, L., Yu, L. and Zhang, Y. (2020) Filtration and Distillation: Enhancing Region Attention for Fine-Grained Visual Categorization. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 11555-11562. https://doi.org/10.1609/aaai.v34i07.6822

[6] Gao, Y., Han, X., Wang, X., Huang, W. and Scott, M. (2020) Channel Interaction Networks for Fine-Grained Image Categorization. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 10818-10825. https://doi.org/10.1609/aaai.v34i07.6712

[7] Zhuang, P., Wang, Y. and Qiao, Y. (2020) Learning Attentive Pairwise Interaction for Fine-Grained Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 13130-13137. https://doi.org/10.1609/aaai.v34i07.7016

[8] Song, J. and Yang, R. (2021) Feature Boosting, Suppression, and Diversification for Fine-Grained Visual Classification. 2021 *International Joint Conference on Neural Networks* (*IJCNN*), Shenzhen, 18-22 July 2021, 1-8. https://doi.org/10.1109/ijcnn52387.2021.9534004

[9] Lin, T., Dollar, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017) Feature Pyramid Networks for Object Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 936-944. https://doi.org/10.1109/cvpr.2017.106

[10] Wang, X., Girshick, R., Gupta, A. and He, K. (2018) Non-Local Neural Networks. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7794-7803. https://doi.org/10.1109/cvpr.2018.00813

[11] Chen, X., Fu, C., Zhao, Y., Zheng, F., Song, J., Ji, R., *et al.* (2020) Salience-Guided Cascaded Suppression Network for Person Re-Identification. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 3297-3307. https://doi.org/10.1109/cvpr42600.2020.00336