2021 IEEE 7th International Conference on Cloud Computing and Intelligent Systems (CCIS) | 978-1-6654-4149-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/CCIS53392.2021.9754631

CFPA-Net: Cross-layer Feature Fusion And Parallel Attention Network For Detection And Classification of Prohibited Items in X-ray Baggage Images

Yifan Wei[1], Yizhuo Wang[1], Hong Song[1*]

[1]School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China
Songhong@bit.edu.cn

Abstract: As objects in the baggage are often heavily overlapped and cluttered, the X-ray baggage inspection is an inherently challenging task. In this paper, we propose a cross-layer feature fusion and parallel attention network named CFPA-Net to detect and classify the prohibited items in X-ray baggage images. The CFPA-Net is based on RetinaNet with three modules: cross-layer feature extraction fusion module (CEF-Module), paralleled attention module (PA-Module) and FreeAnchor. In CEF-Module, an improved feature pyramid network is proposed by adding multi-directional lateral connections for cross-layer feature extraction and fusion. It can help detect objects of various scales and supplement deficient semantic and localization information for low layer and high layer features respectively. PA-Module is presented to learn the feature relationship and fully utilize the extracted features by introducing two paralleled attention subnets Squeeze-and-Excitation module and Non-local module. PA-Module can help improve the performance of detecting and classification by emphasizing useful features, suppressing useless features selectively and capturing long-range dependencies in images. FreeAnchor is adopted to deal with the restriction of hand-crafted anchor assignment according to Intersection-over-Unit. It can help find the best anchor for each object by learning, and improve the performance of detecting slender objects and the ones in crowded scenes. On the public dataset OPIXray, CFPA-Net achieves 85.82% detection mean Average Precision. Moreover, achieving 81.61% classification mean Average Precision on the SIXray10 dataset. The experimental results show that our proposed CFPA-Net is more accurate and robust for the X-ray baggage inspection with densely occluded objects and complicated backgrounds.

Keywords: Prohibited items; X-ray baggage inspection; Cross-layer feature fusion; Paralleled attention subnets

1 Introduction

Security inspection with X-ray scanners is widely used. However, in the real-world scenario, objects in baggage are heavily overlapped and occluded with each other, and after long-time watching amounts of the complex X-ray images, security inspectors are fatigued, which may cause bad consequences to the public. Therefore, automatic X-ray image detection and classification remain a key issue in security inspection.

Deep learning has great achievement for object detection and classification in recent years. Miao et al. [1] propose a model named CHR for classifying the prohibited items from the SIXray dataset. The researchers in [2, 3] both proposed the method that each object is extracted by iteratively picking contour-based transitional information and a single feed-forward convolutional neural network is used for the recognition. Wei et al. [4] proposed the De-occlusion Attention Module (DOAM). This module simultaneously leverages edge information and material information of the prohibited items to generate the attention maps and feature maps for detection. In general, the previous methods based on deep learning neglect the shortage of semantic information in low level features and localization information in the topmost level features. And these methods generally consider that all features are equally important for detection and classification tasks. Moreover, these methods are not conducive to the performance of detection and classification under the scene of serious occlusion.

To address these issues, we propose a cross-layer feature fusion and parallel attention network named CFPA-Net for classification and detection of prohibited items, which includes a cross-layer feature extraction fusion module (CEF-Module), a paralleled attention module (PA-Module) and FreeAnchor. The CEF-Module is designed to obtain balanced and integrated features. And the PA-Module is introduced to capture long-range contextual information, which further refines features to be more discriminative. FreeAnchor is employed to select more suitable anchors for slender objects or objects in crowded scenes. Compared with other methods for X-ray baggage inspection and classification, our method has great advantages in extracting and integrating interior object features for X-ray baggage images detection and classification. The contributions of our work are as follows:

• We propose the CEF-Module by combining multi-directional and cross-layer lateral connections with the original feature pyramid network (FPN) [5].This module can supplement deficient semantic and localization information for low layer and high layer features respectively for improving the feature extraction

978-1-6654-4149-0/21/$31.00 ©2021 IEEE

capabilities.

• We propose the PA-Module by employing Squeeze-and-Excitation (SE) module [6] and Non-local module [7] to get paralleled attention subnets. The PA-Module is used to get more discriminative features by emphasizing task-related objects information and suppressing less useful ones from the channel and spatial aspects.

• We evaluate our framework on two public X-ray datasets (OPIXray and SIXray10) and compare it with other methods. The experiments show that our proposed CFPA-Net outperforms the state-of-the-art methods for the detection and classification task of prohibited items under the scene of serious occlusion and complex background.

2 Method

The architecture of our proposed CFPA-Net is shown in Figure 1. The input of CFPA-Net is an X-ray baggage image. The goal of CFPA-Net is to recognize the categories of prohibited items in the input image and locate the objects in the output.

2.1 CFPA-Net architecture

CFPA-Net is proposed by improving the RetinaNet [8], which is a representative architecture of single-stage detection approaches. RetinaNet can be divided into feature pyramid network (FPN) and two task-specific subnetworks. Since low layers and high layers lack semantic information and position information respectively in the original FPN network, we propose a cross-layer feature extraction fusion module (CEF-Module) based on FPN. For getting more discriminative features and emphasizing task-related objects information, we add a paralleled attention module (PA-Module) after CEF-Module. Moreover, we employ FreeAnchor in two subnetworks of RetinaNet for solving the problem in matching anchors with hand-crafted Intersection-over-Union (IoU) criterion.

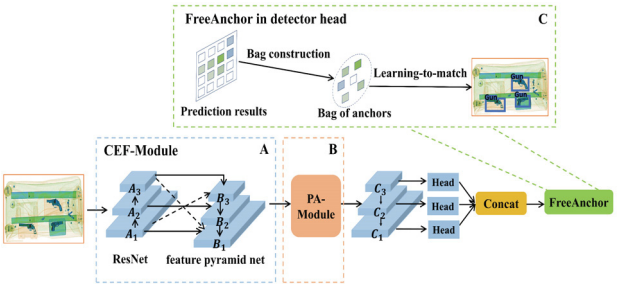

Figure 1 The architecture of CFPA-Net. Part A represents CEF-Module, Part B represents PA-Module, and Part C represents FreeAnchor in detector head.

Figure 1 shows the details of the architecture of the proposed CFPA-Net, including CEF-Module (Part A of Figure 1), PA-Module (Part B of Figure 1) and FreeAnchor in detector head (Part C of Figure 1).

2.2 Feature extraction based on multi-directional lateral connections and cross-layer fusion

In FPN, it has been found that high-level features contain much semantic information that is beneficial for classification. Furthermore, low-level features have much detailed and spatial information, which is beneficial for location. Object detection requires both accurate positions and accurate categories, therefore cross-layer fusion is needed.

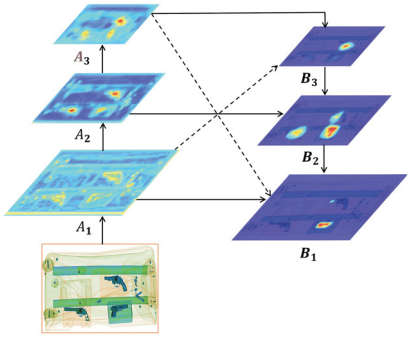

Figure 2 CEF-Module. As illustrated, multi-layer feature maps are generated by cross-layer and multi-directional lateral connections.

Figure 2 is the proposed cross-layer feature extraction fusion module (CEF-Module). It includes a bottom-top path and a top-down one. The whole module is based on FPN with the backbone of Resnet-50. We denote the outputs of Resnet as $\{A_1,\ A_2,\ A_3\}$. The final multi-level feature maps are called $\{B_1,\ B_2,\ B_3\}$, corresponding to $\{A_1,\ A_2,\ A_3\}$. Based on FPN, by using bilinear interpolation, we up-sample the highest pyramid level A_3, making its size the same as the lowest pyramid level A_1, and down-sample the lowest pyramid level A_1 making its size the same as the highest pyramid level A_3. Then, we attach a 1×1 convolutional layer on the feature maps which has been down-sampled or up-sampled to produce the intermediate feature maps. These intermediate feature maps are then merged with the feature maps that the 1×1 convolutional layer is attached to the original A_1, A_3 by element-wise addition. This process can generate the final lateral connections of A_1, A_3. The $B_i - th$ level is defined as:

$$B_i = \begin{cases} B_{i+1} \oplus A_i \oplus A_{i+2} & ,i = 1 \\ B_{i+1} \oplus A_i, & ,i = 2 \\ A_{i-2} \oplus A_i, & ,i = 3 \end{cases} \quad (1)$$

where $\oplus$ is an addition operation.

We add a 1×1 convolutional layer on the level A_2 to generate the lateral connection of A_2. Then the lateral connections of A_1, A_2, A_3 merge feature maps of the same spatial size from the top-down pathway by element-wise addition to produce the fused level B_i. In the up-sampling process of the feature pyramid network, there is a long path from the topmost feature to the low-level structure, leading to information loss. It is unfavorable for the subsequent fusion of feature maps,

due to the deficiency of rich semantic information in low levels and localization information in the topmost levels. Therefore, before constructing a top-down pyramid, we supplement the original spatial information and semantic information corresponding to high-level and low-level feature maps. The feature maps producing by CEF-Module are utilized to detect objects of various scales.

2.3 Attention module with parallel attention subnets

Low-level and high-level information is complementary for object detection and classification. We resize the level features $\{B_1,\ B_3\}$ to the same size as the middle level B_2, with max-pooling and interpolation respectively, as shown in Figure 3. This process aims to integrate multi-level features information from each resolution and preserve the features semantic hierarchy at the same time. The resized level features are named $\{P_1, P_2, P_3\}$, then the fused features are obtained by addition and averaging as:

$$P = \frac{1}{L}\sum_{l=l_{min}}^{l_{max}} P_l \tag{2}$$

where P represents the final fused features, P_l are the features at resolution level l. L is the number of multi-level features, and the index of lowest and highest levels are l_{min} and l_{max} respectively.

We use two parallel modules, SE module and Non-local module, as the channel attention network and the location attention network respectively, for enhancing the features P.

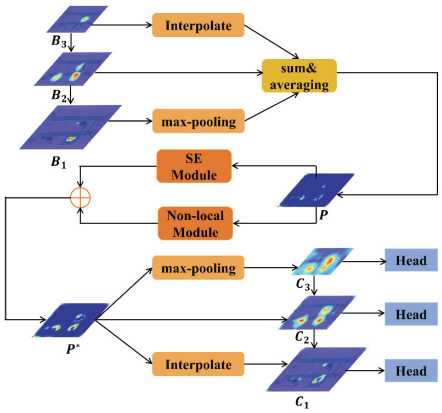

Figure 3 PA-Module. The parallel attention subnets are to refine the feature maps.

Specifically, the process is utilized to strengthen the features P from the channel level by explicitly modeling interdependencies between channels and learning the important degree of each channel. It aims to emphasize useful features and suppress useless features selectively. The features P can be enhanced from the location level by capturing long-range and nonadjacent dependencies for global information. Then, the enhanced features captured by the two attention networks are merged to generate features P^* by element-wise addition. The obtained features P^* are rescaled to the sizes of original levels $\{B_1,\ B_3\}$ using interpolation and max-pooling procedure respectively. Finally, we add the rescaled layers and the initial levels $\{B_1,\ B_2,\ B_3\}$ for generating the levels $\{C_1,\ C_2,\ C_3\}$, aiming to decrease the information loss and alleviate gradient disappearance in backpropagation.

2.4 FreeAnchor in the detector head

In the X-ray Baggage Images, on the one hand, the center of many slender objects such as scissors or wrenches is not close to the most representative features. This induces that the anchors aligned with ground truth boxes might correspond to fewer representative features or most background. On the other hand, objects in baggage are crowded and occluded with each other, so it is infeasible to match proper anchors for objects using IoU. To deal with these issues, FreeAnchor [9] was adopt to find the best anchor for each object. The process of FreeAnchor can be visualized in Figure 1.

3 Experimental results and analysis

We will evaluate our proposed method on two public X-ray datasets in the section. We first show the datasets and then introduce our evaluation metrics and implementation details. Finally, we compare the results with other models.

3.1 Datasets

OPIXray [4] is a public dataset for the detection of prohibited items, which contains 8885 colored X-ray images of 5 categories, namely, Folding Knife, Straight Knife, Scissor, Utility Knife, and Multi-tool Knife. The dataset is partitioned into a training set (7109 images) and a testing set (1776 images), and the ratio of training set and testing set is about 4:1.

SIXray [1] is a dataset containing 1,059,231 colored X-ray images. Out of these images, 8,929 images contain the prohibited items, and the others contain only the normal baggage items. The dataset contains six categories of prohibited items, namely, hammer, scissors, pliers, wrench, knife, and gun. Furthermore, the dataset is divided into three subsets namely SIXray10, SIXray100, and SIXray1000, and the hammer class with 60 samples is not contained in three subsets. Each subset is further arranged into a training set and a test set, the rate of training/testing images is 4:1, only SIXray10 is used in our experiments.

3.2 Evaluation metrics

We will evaluate the performance according to the following metrics:

Detection mean Average Precision: To calculate mean Average Precision (mAP), we first sort the bounding boxes according to their confidence scores. Then, we calculate IoU for each bounding box. We can calculate the mAP scores using the IoU threshold of 0.5.

$$mAP = \frac{1}{C}\sum_{n=1}^{C} AP(n) \tag{3}$$

where C denotes the number of prohibited item categories. Finally, we find mAP by averaging AP values.

Classification mean Average Precision: For image classification, we apply the evaluation metric in the PascalVOC image classification task [10] Specifically, the test dataset is sorted according to the confidence containing the specified object, then the mean Average Precision can be calculated.

3.3 Implementation details

Our networks were implemented on PyTorch. We use stochastic gradient descent as the optimization method and the learning rate of our network is 0.005. The batch size of our network is 16, and our momentum and weight decay are 0.9 and 0.0001 respectively. Finally, we evaluate the mAP of the object detection on the OPIXray dataset. Then we evaluate the classification mean Average Precision on the SIXray10 dataset. We train the networks using 4 NVIDIA 2080Ti GPUs; for testing, 1 NVIDIA 2080Ti GPU is required.

3.4 Evaluations on OPIXray Dataset for Detection

First, we use the OPIXray dataset for training and evaluating our model. We show the performance of different methods on the OPIXray dataset and offer some further analysis. Shown in Table I, our model achieves 85.82% in terms of mAP. The result shows that our model outperforms previous methods which are tested on this dataset. Moreover, our proposed framework outperforms SSD [11], SSD+DOAM [4], FCOS [12], FCOS+DOAM [4], TST [13], and Faster R-CNN [14] by 14.93%, 11.81%, 3.8%, 3.41%, 10.5%, and 0.36% respectively. We consider that the reason is that our model can better improve the feature extraction capabilities and learn the feature relationship from the channel and space aspects. In contrast, other models do not obtain the balanced features and ensure no massive loss of information during up-sampling. And they neglect the shortage of semantic and localization information in low level features and high level features respectively. Our method is more stable in dealing with the situation of cluttered backgrounds and occluded objects. Figure 4 shows some detection results using our proposed method on the OPIXray dataset.

Table I Average Precision (%) and mean Average Precision (%) for detection on the testing set of the OPIXray dataset. FO, ST, SC, UT and MU are the abbreviations of Folding Knife, Straight Knife, Scissor, Utility Knife and Multi-tool Knife, respectively.

Method	mAP	FO	ST	SC	UT	MU
SSD	70.89	76.91	35.02	93.41	65.87	83.27
SSD+DOAM	74.01	81.37	41.50	**95.12**	68.21	83.83
FCOS	82.02	86.41	68.47	90.22	78.39	86.60
FCOS+DOAM	82.41	86.71	68.58	90.23	78.84	87.67
TST	75.32	80.24	56.13	89.34	72.89	78.02
Faster R-CNN	85.46	86.20	75.53	90.19	**86.49**	88.88
Our method	**85.82**	**87.66**	**76.05**	90.53	85.91	**88.94**

3.5 Ablation study

To evaluate the effectiveness of different modules, we conduct an ablation study for analyzing CEF-Module, PA-Module, and FreeAnchor. In the experiments, RetinaNet is selected as a foundation of the proposed method. V1 represents RetinaNet with CEF-Module; V2 represents RetinaNet with PA-Module; V3 represents RetinaNet with FreeAnchor; V1+V2 represents RetinaNet with CEF-Module and PA-Module; V1+V3 represents RetinaNet with CEF-Module and FreeAnchor; V2+V3 represents RetinaNet with PA-Module and FreeAnchor. our method uses CEF-Module, PA-Module and FreeAnchor based on RetinaNet. As Table Ⅱ shown, for the OPIXray dataset, that the CEF-Module, PA-Module and FreeAnchor improved the performance by 1.22%, 1.72% and 1.01% respectively compared with the RetinaNet network. Moreover, the PA-Module's improvement is 2.3% more than V1. Furthermore, FreeAnchor's improvement is 0.52% than V1+V2. Furthermore, our method's improvement is 4.04% than RetinaNet. The comparison results verify the effectiveness of the CEF-Module and PA-Module. Specifically, the CEF-Module supplements deficient semantic and localization information for low layer and high layer features respectively. And the PA-Module can get more discriminative features and emphasize task-related objects information. Furthermore, the results reveal that features can be more balanced and abundant by cross-layer fusion. And the attention of different aspects is essential for making the model more focus on the relative information about the current task. FreeAnchor can also contribute to the final results, proving that it is favorable to select anchor for the object by learning.

Table Ⅱ Average Precision (%) and mean Average Precision (%) for detection results of ablation study on OPIXray.

Method	mAP	FO	ST	SC	UT	MU
RetinaNet	81.78	87.75	66.82	89.29	76.93	88.08
V1	83.00	86.34	67.61	90.08	83.54	87.44
V2	83.50	86.82	67.38	90.25	85.50	87.54
V3	82.79	87.79	69.49	90.00	78.74	87.92
V1+V2	85.30	88.12	**76.77**	90.32	82.97	88.30
V1+V3	83.70	**88.66**	67.57	89.83	84.73	87.70
V2+V3	85.72	88.40	75.55	90.16	**86.52**	87.96
Our method	**85.82**	87.66	76.05	**90.53**	85.91	**88.94**

3.6 Evaluations on SIXray10 Dataset for classification

We also use the dataset SIXray10 to train and evaluate the classification performance of our method. As shown in Table Ⅲ, our method achieves 81.61% in terms of classification mean Average Precision, which outperforms the tested methods on the dataset. It is evident from Table Ⅲ that the proposed method performs better especially for Wrench, Scissors. However, the performance on Gun, Knife, and Pliers is not the best. We consider that this is because it is trained and tested on an imbalanced ratio of positive and negative samples. In the test dataset, the model classifies the normal baggage content as the prohibited category incorrectly. Moreover, our proposed framework outperforms ResNet34 [15], ResNet34+CHR [1], ResNet50 [15], ResNet50+CHR [1], ResNet101 [15],

ResNet101+CHR [1], Inception-v3 [16], Inception-v3+CHR [1], DenseNet [17], and DenseNet+CHR [1] by 6.78%, 4.41%, 4.76%, 3.67%, 4.23%, 2.24%, 4.6%, 2.12%, 4.25%, 2.05%, respectively. According to the results, we can observe how the proposed method effectively classifies the samples from the scenes of clutter and occlusion on an imbalanced set. The result is that by using CEF-Module and PA-Module we can get more discriminative features, supplement the semantic information for low-level feature maps and improve the feature extraction capabilities.

Table Ⅲ Average Precision (%) and mean Average Precision (%) for classification on the testing set of the SIXray10 dataset. * represents current model adding CHR module.

Method	Mean	Gun	Knife	Wrench	Pliers	Scissors
ResNet34	74.83	89.71	85.46	62.48	83.50	52.99
ResNet34*	77.20	87.16	87.17	64.31	85.79	61.58
ResNet50	76.85	**90.64**	**87.82**	63.62	84.80	57.35
ResNet50*	77.94	87.55	86.38	69.12	85.72	60.91
ResNet101	77.38	87.65	84.26	69.33	85.29	60.39
ResNet101*	79.37	85.45	87.21	71.23	88.28	64.68
Inception-v3	77.01	90.05	83.80	68.11	84.45	58.66
Inception-v3*	79.49	88.90	87.23	69.47	86.37	65.50
DenseNet	77.36	87.36	87.71	64.15	87.63	59.95
DenseNet*	79.56	87.05	85.89	70.47	**88.34**	66.07
Our method	**81.61**	86.07	86.33	**72.44**	87.28	**75.95**

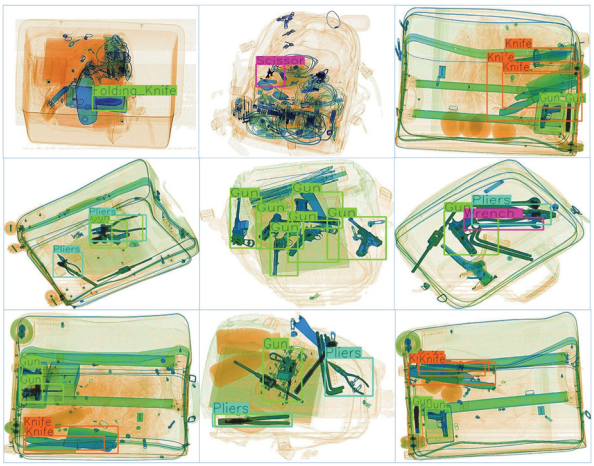

Figure 4 Some detection results using our method on the OPIXray test dataset and the SIXray10 test dataset.

4 Conclusions

In this work, we developed a new framework based on RetinaNet with the cross-layer feature extraction fusion module (CEF-Module) and paralleled attention module (PA-Module) for X-ray baggage inspection. In addition, FreeAnchor was also considered to address the issue of matching anchors for objects and improve detection accuracy. We evaluate our proposed framework on the OPIXray dataset and the SIXray10 dataset. The experimental results show that our method achieves better performance than other state-of-the-art models. Moreover, it is more effective for detecting and classifying occluded and cluttered objects, demonstrating its potential application in the real world.

References

[1] Miao C, Xie L, Wan F, et al. Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2119-2128.

[2] Hassan T, Khan S H, Akcay S, et al. Cascaded Structure Tensor Framework for Robust Identification of Heavily Occluded Baggage Items from Multi-Vendor X-ray Scans. arXiv preprint arXiv:1912.04251, 2019.

[3] Hassan T, Khan S H, Akcay S, et al. Deep CMST Framework for the Autonomous Recognition of Heavily Occluded and Cluttered Baggage Items from Multivendor Security Radiographs. CoRR, dec, 2019, 14: 17.

[4] Wei Y, Tao R, Wu Z, et al. Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. Proceedings of the 28th ACM International Conference on Multimedia. 2020: 138-146.

[5] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[6] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[7] Wang X, Girshick R, Gupta A, et al. Non-local neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.

[8] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[9] Zhang X, Wan F, Liu C, et al. Freeanchor: Learning to match anchors for visual object detection. arXiv preprint arXiv:1909.02466, 2019.

[10] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge. International journal of computer vision, 2010, 88(2): 303-338.

[11] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector. European conference on computer vision. Springer, Cham, 2016: 21-37.

[12] Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional one-stage object detection. Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9627-9636.

[13] Hassan T, Werghi N. Trainable Structure Tensors for Autonomous Baggage Threat Detection Under Extreme Occlusion. Proceedings of the Asian Conference on Computer Vision. 2020.

[14] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 2015, 28: 91-99.

[15] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[16] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.

[17] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-470.