Dual Channel Attention Networks

Jingchen Bian* & Yugui Liu
University of Chinese Academy of Sciences, Beijing, China

*E-mail: bianjingchen17@mails.ucas.ac.cn

Abstract. Channel attention is currently widely used in Computer Vision. Most existing channel attention networks are proposed based on Squeeze-and-Excitation Networks (SE-Net), which can obtain excellent performance by designing complex structures, however, they also have more additional network parameters and higher floating point operations per second (FLOPs). We propose a novel lightweight attention structure called Dual Channel Attention Networks (DCA-Net). By introducing the channel attention preprocessing module and using 1-D convolution with K=1, DCA-Net has a more straightforward and delicate structure. We add DCA-Net to ResNet and perform image classification experiments on CIFAR-100 dataset, object detection and instance segmentation experiments on MS-COCO dataset. Experimental results show that our DCA-Net achieves better results than the existing attention networks, such as SE-Net, ECA-Net and so on. For example, in the image classification task on CIFAR-100, the parameter amount of DCA-Net using ResNet-50 decreases by 50.03% compared to SE-Net using ResNet-101, and decreases by 44.48% compared to ECA-Net using ResNet-101, GFLOPs decreased by 48.41% and 48.21%, respectively. At the same time, the Top-1 accuracy of DCA-Net using ResNet-50 is 0.3% higher than that of SE-Net using ResNet-101, and 0.86% higher than that of ECA-Net using ResNet-101.

1. Introduction

In 2012, Krizhevsky et al. used deep convolutional neural networks (CNN) called AlexNet [1] and won first place in the classification competition in the ImageNet Large Scale Visual Recognition Challenge (LSVRC). After that, more and more deep network structures [2][3][4][5][6][7] were proposed to obtain better performance. Recently, adding attention module, especially channel attention module to deep convolutional neural network [8][9][10][11] has been widely concerned by researchers. Attention modules in deep learning was first proposed in machine translation task [12], and has achieved significant results. Later, in the field of computer vision, attention module has also been widely used to improve network performance. The basic principle of the attention modules in deep learning is easy to understand. In essence, it believes that each layer in the network (maybe from different channels, maybe from different regions) has different characteristics, and the later layers should pay more attention to the vital information and suppress the unnecessary information.

Channel attention is a type of attention structure used in computer vision. The most representative channel attention module is the Squeeze-and-Excitation Network (SE-Net)[8]. SE-Net focuses on the relationship between feature channels and adds an attention mechanism to the feature channels. By introducing the Squeeze module and the Excitation module, SE-Net automatically obtains the importance of all feature channels through learning and uses the obtained importance to increase or suppress features that are vital or unnecessary for the current task. Some attention modules are proposed later, like CBAM[9], A²-Nets[10], and etc., which improve performance by designing...
complex attention structures. Besides, ECA-Net[11] proposed a simple attention structure that creatively replaces the fully connected layer in SE-Net with 1-D convolution, which can avoid dimensionality reduction while reducing the number of parameters and FLOPs.

By analyzing SE-Net and ECA-Net, we found that compared with SE-Net, ECA-Net does bring about performance improvements by avoiding dimensionality reduction. However, on the other hand, ECA-Net only considers the information exchange between the current channel and its k neighborhood channels, thereby losing the interaction relationship of all channels that SE-Net has. We believe this is a performance bottleneck of ECA-Net.

To solve the above problem of ECA-Net, we propose a novel attention network called Dual Channel Attention Networks (DCA-Net). The structure of DCA-Net is shown in Figure 1. Compared with ECA-Net, the most significant difference of DCA-Net is the addition of a pre-attention module between Adaptive AvgPool2d and 1-D convolution. By performing a simple matrix transposition operation and a softmax operation, the pre-attention module can learn the one-to-one and one-to-many interaction relationships between channels without increasing parameters. Then 1-D convolution is conducted to get the many-to-many interaction between the channels. It is worth mentioning that in DCA-Net, the convolution kernel size of 1-D convolution is set to 1 or 3 to obtain the interaction relationships between all channels, which can improve the network performance.

Our contributions are as follows:
(1) By analyzing SE-Net and ECA-Net, we found that getting the interaction between all channels is very important for channel attention.
(2) We propose a novel lightweight attention network called DCA-Net. DCA-Net uses a pre-attention module mainly composed of matrix operations and a post-attention module mainly composed of 1-D convolution layers, which can easily obtain interaction relationships between all channels. Depends on the simple structure, after adding it to a deep neural network such as ResNet[4], DCA-Net has fewer network parameters compared with other attention networks.
(3) Experimental results on CIFAR-100[13] and MS-COCO[14] datasets show that DCA-Net has more excellent performance than SE-Net and ECA-Net, and which is proved that DCA-Net is suitable for tasks such as image classification, object detection and instance segmentation.

![Figure 1. The overall structure of DCA-Net when deployed in ResNet.](image)

2. Related Works

2.1. Squeeze-and-Excitation Networks
In 2017, Hu et al. proposed a structure using attention mechanism in the channel-wise channel called Squeeze-and-Excitation Networks (SE-Net)[8], which won the first place in the 2017 ImageNet classification task. The SE-Net can be easily implanted into a backbone CNN such as ResNet[4], MobileNet. For the feature map output from the convolutional layer, SE-Net uses two fully connected layers to process, and obtains a one-dimensional vector with the same number of channels as the weight score of each channel, and then applies the score to the corresponding channel respectively. Experimental results show that SE-Net is useful for image classification. When a classification network using the SE-Net structure as a backbone, the performance of object detection and instance segmentation can also be improved.
2.2. Dual Attention Network
In 2018, Fu et al. proposed Dual Attention Network (DAN)[15]. DAN is a network structure dedicated to Scene Segmentation. After the backbone CNN is put into the two parallel Attention Modules, the finally extracted features are significantly improved. Among them, DAN’s Channel Attention Module puts the features with the size C*H*W extracted by the backbone CNN into the channel attention module, and obtains 1 to N’s attention through matrix transposition operation.

2.3. Efficient Channel Attention Network
In 2020, Wang et al. proposed the Efficient Channel Attention Network (ECA-Net)[11] based on SE-Net by using one-dimensional convolution to replace the full connection to avoid dimensionality reduction, which dramatically reduces network parameters. ECA-Net uses 1-D convolution to capture local cross-channel interactions, so the kernel size of the convolution determines the coverage of cross-channel interactions. Its choice of the convolution kernel is determined using the classic exponential function formula related to the number of channels. In ResNet-50[4] and ResNet-101[4], the maximum size of channels is generally 2048 or much smaller, so the convolution kernel size of ECA-Net is generally set to 3 or 5 when applied to ResNet. Therefore, the coverage of the cross-channel interaction captured by ECA-Net is only 3 or 5, but SE-Net obtains the interaction of all channels through two fully connected layers. Specifically, the two fully connected layers in SE-Net allow the network to learn many-to-one and one-to-many weight relationships, but ECA-Net only learns many(3 or 5)-to-one weight relationships.

Based on them, we propose a novel Dual Channel Attention Network (DCA-Net). DCA-Net consists of a pre-attention module and a post-attention module. Through the combination of the two modules, DCA-Net can obtain the weight relationships between all channels while avoiding dimensionality reduction, while the amount of additional parameters is minimal.

3. Our Approach

3.1. Motivation
By exploring SE-Net[8] and ECA-Net[11], we found that the most significant difference between the two networks is that ECA-Net replaces the two fully connected layers with a simple one-dimensional convolution, which dramatically saves the number of parameters and avoids dimension reduction, and achieves competitive results.

However, ECA-Net uses kernel size like 3 or 5 depending on a prescribed formula which will form a too small scope of multichannel interaction relationships; correspondingly, SE-Net obtains the interaction relationships between all channels. The difference in scope of interaction relationships can cause a performance bottleneck of ECA-Net.

We first conduct experiments by expanding the equivalent receptive field and test the effect of directly increasing the receptive field of 1-D convolutional on ECA-Net. After that, we make further improvements to ECA-Net. By designing the attention network as a pre-attention module and a post-attention module, our structure obtains 1 to N (N refers to the total number of channels) relationship between the channels through the pre-attention module, and then obtains the N to N relationship to get a global multi-channel relationship while avoiding dimension reduction. Our proposed structure is named DCA-Net.

3.2. Pre-attention Module
ECA-Net[11] performed different tests on the kernel size of the 1-D convolution in the experiments. Experimental results show that the performance of ECA-Net at K = 5 and K = 9 is significantly better than K=3 on ResNet-50[4] and ResNet-101[4]. Although ECA-Net finally chose K = 3 as the default setting to save parameters, we think that this proves the necessity of acquiring more interactions between channels for the channel attention mechanism. Based on the above research, we propose a
pre-attention module, which allows the network to finally obtain the interaction between all channels without increasing too many parameters.

The structure of pre-attention module is shown in Figure 2. Our pre-attention module is inspired by Dual Attention Network for Scene Segmentation (DAN)[15] proposed by Fu et al. Since the original DAN occupied too much memory capacity, it can only be added to the head of the overall network structure. Our network first converts the C*H*W feature map into C*1*1 through Adaptive AvgPool2D, and then performs other operations. Thanks to this design, our module dramatically reduces the memory capacity compared to DAN and can be easily placed in each block of ResNet.

In each block of ResNet, we first use Adaptive AvgPool2D to convert the feature map of size C*H*W into F ∈ R^{C*1*1} after all convolutional layers, and then enter the pre-attention module. We transform the original 1*1 feature map flatten into a second-order tensor to form a matrix of E ∈ R^{C*1}. Then, we transpose this matrix to form a matrix E^T ∈ R^{1*C}. We multiply these two matrices, and use softmax later. In this way, we obtain an adjacency matrix A ∈ R^{C*C}, which is formulated in Equation 1.

\[
A = \text{Softmax}(E \cdot E^T)
\]  

(1)

Each cell in the adjacency matrix A can represent a 1 to 1 attention relationship between different channels in the original feature map. We multiply the adjacency matrix A and E, reshape the results to B ∈ R^{C*1*1}.

\[
O = \alpha \cdot B + F
\]  

(2)

Then multiply B by a weight α that is gradually learned from 0, and add it to the input of pre-attention module F and get O, which is the input of post-attention module. The operation is formulated in Equation 2. The purpose of α is to prevent the value of many cells from changing too small later and making information loss.

Finally, through the pre-attention module, we can obtain the 1 to N interaction relationships between all channels.

Figure 2. The structure of pre-attention module.

3.3. Post-attention Module

Based on the pre-attention module, the input of the post-attention module has already obtained the 1 to N interaction relationships between different channels, so we think that in the post-processing module, only using 1-D convolution with kernel size=1 or 3 could easily obtain the global channel interaction relationships of N to N.

The structure of the post-attention module is shown in Figure 3, which is very similar to that of ECA-Net [11]. The most significant difference between two structure is the existence of the pre-attention module, and our post-attention module directly uses 1-D convolution with a convolution kernel = 1, which further reduces the parameters compared to ECA-Net.

We found that as the kernel size of 1-D convolution layers increasing, the performance of DCA-Net has not improved significantly. We set the kernel size of 1-D convolution layers to 1 by default in experiments to save parameters.
3.4. Usage in ResNet

In Figure 1, we show how to deploy DCA-Net in ResNet[4]. We set the pre-attention module and post-attention module after all the convolution layers in each block of ResNet, respectively. The structure of DCA-Net is straightforward, and the parameters we add to ResNet are mainly from 1-D convolutions with kernel size of 1, which makes parameter amount less than that of ECA-Net and SE-Net. Furthermore, DCA-Net can be easily used in other CNNs, such as VGG[3], DenseNet[5], ResNeXt[16] and so on.

4. Experiments

4.1. Datasets

Our experiments are based on CIFAR-100[13] for image classification, MS-COCO[17] for object detection and instance segmentation. CIFAR-100 has 100 categories, and each category includes 500 training images and 100 test images, we use Train split (50k images) for training and report the performance on Val split (10k images). MS-COCO contains 80 categories and more than 1.5 million object instances. We use the Train split (115K images) for training and report the performances on Val split (5K images).

4.2. Evaluation Metrics

To evaluate the performance of DCA-Net, we consider multiple evaluation metrics. We used Top-1 accuracy and Top-5 accuracy on Val split for image classification experiments on CIFAR-100 dataset. For object detection and instance segmentation experiments on MS-COCO dataset, we used AP metrics to evaluate the performance. AP refers to average precision rate and is averaged over categories and IoU. AP can be used to evaluate the performance under different object scales, including small objects (area < 32²), medium objects (32² < area < 96²) and large objects (area < 96²).

4.3. Training Details

To evaluate the performance of DCA-Net on image classification, we employ two widely used CNNs as backbone networks, including ResNet-50[4], ResNet-101[4]. To make the network adapt to 32 * 32 input image size of CIFAR-100, we reduce the convolution kernel of the first layer of the original ResNet convolution from 7 to 3, and other structures are the same as the original ResNet. During training, we use a random crop of size 32 and a random horizontal flip for image enhancement. We follow the hyperparameter settings in Improved Regularization of Convolutional Neural Networks with Cutout[18], in which learning rate is initialized to 0.1 and divided by 5 at 60th, 120th and 160th epochs, train for 200 epochs with batch size of 128, weight decay of 5e-4, and momentum of 0.9.

We also evaluate DCA-Net on MS-COCO using RetinaNet[19] and Mask R-CNN[20] along with Feature Pyramid Networks(FPN)[21]. ResNet-50 is used as the backbone network of our method. We implement all detectors by MMDetection toolkit[17] and employ the default settings as ECA-Net. Specifically, the input images are resized to 1333*800, and then all models are optimized using SGD with weight decay of 1e-4, momentum of 0.9, and mini-batch size of 8 (4 GPUs with 2 images per GPU). The learning rate is initialized to 0.01 and divided by 10 at 9th, 12th epochs. We train all detectors within 12 epochs on train2017 of MS-COCO and report the results on val2017 for.
comparison. All experiments are run on a Linux Server with 4 Tesla P100 GPUs and an Intel(R) Xeon Silver E5-2640 CPU@2.40GHz. We implement all models with PyTorch toolkit.

4.4. Image Classification Experiments on CIFAR-100

4.4.1. Effect of Different Kernel Sizes \((k)\) in Post-attention Module. Since the post-attention module uses the same 1-D convolution layers as ECA-Net, our DCA-Net will also involve the selection of kernel size. In this subsection, we test the impact of different kernel size of 1-D convolution layers on the performance of DCA-Net through experiments. Specifically, we use ResNet-50 as the backbone network, set the kernel size to 1, 3, 5, and 7, respectively, and then conduct an image classification experiment with DCA-Net on CIFAR-100. Experimental results in Table 1 show that DCA-Net achieves the best performance when kernel size is 3. The Top-1 accuracy of Kernel size 5, 7 are not as good as that of kernel size 1. It depends existence of the pre-attention module that enables the network to acquire the 1 to N interaction relationships between all channels. Unlike ECA-Net, kernel size of 1-D convolution layers in DCA-Net does not directly affect the scope of the network to obtain interaction relationships between different channels, which is also the most significant advantage of DCA-Net compared to ECA-Net.

| Kernel Size | Top-1 | Top-5 |
|-------------|-------|-------|
| 1           | 80.97 | 95.51 |
| 3           | 81.24 | 95.69 |
| 5           | 80.54 | 95.12 |
| 7           | 79.68 | 94.97 |

4.4.2. Comparison Using CNNs. We compared ECA-Net with other state-of-the-art channel attention methods on ResNet-50 and ResNet-101. Evaluation metrics include network parameters, floating point operations per second (FLOPs), Top-1 and Top-5 accuracy. For a fair comparison, we use open source code to test them on the same computing platform. The results are shown in the Table 2. Experimental results show that DCA-Net has no significant increase in network parameters and FLOPs compared to ResNets and ECA-Net using ResNets, but has a significant improvement in performance. Please note that the Top-1 accuracy of DCA-Net using ResNet-50 has greatly exceeded the Top-1 accuracy of SE-Net and ECA-Net using ResNet-101. At the same time, the parameter amount of DCA-Net using ResNet-50 reduces by 50.03% compared to SE-Net using ResNet-101 and decreases 44.48% compared to ECA-Net using ResNet-101 and GFLOPs decrease by 48.41% and 48.21%, respectively. Experimental results show that compared to SE-Net and ECA-Net, DCA-Net has higher performance and efficiency in image classification on CIFAR-100.

| CNN Model    | Param. | GFLOPs | Top-1 | Top-5 |
|--------------|--------|--------|-------|-------|
| ResNet-50    | 22.607M| 1.215  | 79.43 | 95.18 |
| +SE block    | 25.006M| 1.222  | 80.34 | 95.16 |
| +ECA block   | 22.607M| 1.217  | 80.24 | 95.69 |
| +DCA(Ours)   | 22.607M| 1.217  | 81.24 | 95.69 |
| ResNet-101   | 40.719M| 2.348  | 80.19 | 95.3  |
| +SE block    | 45.243M| 2.359  | 80.94 | 95.28 |
| +ECA block   | 40.720M| 2.350  | 80.38 | 95.12 |
| +DCA(Ours)   | 40.720M| 2.350  | 81.28 | 95.72 |
4.5. Object Detection Experiments on MS-COCO
In this subsection, we evaluate DCA-Net on object detection task using Faster R-CNN [22], Mask R-CNN [20] and RetinaNet [19]. We mainly compare DCA-Net with ResNet[4], SE-Net[8] and ECA-Net[11].

4.5.1. Comparisons Using RetinaNet. We first use one-stage detector RetinaNet to verify the performance of DCA-Net in object detection. As shown in Table 3, the AP of DCA-Net on ResNet-50 is 1.2% higher than the original ResNet-50, and 0.2% higher than that of the ECA-Net AP based on ResNet-50. At the same time, there is no significant difference from ECA-Net in parameters and GFLOPs. DCA-Net is superior to the original ResNet, SE-Net and ECA-Net in the detection performance of small objects. As we all know, the detection of small objects has always been the difficulty of object detection.

4.5.2. Comparisons Using Mask R-CNN. Then, we use the two-stage detector Mask R-CNN to further verify the performance of DCA-Net. As shown in Table 3, DCA-Net get the same AP as ECA-Net, specifically, the AP of DCA-Net based on ResNet-50 is 0.9% higher than the original ResNet, which is in line with the performance of ECA-Net, but in terms of the detection performance of small objects, DCA-Net is 0.6% higher than ECA-Net, which shows that DCA-Net is more convenient to obtain the characteristics of small objects, and which also reflects the versatility of DCA-Net in the field of object detection.

Table 3. Object detection results of different methods on COCO val2017.

| Detectors | Methods     | Param. | GFLOPs | AP | AP50 | AP75 | APs | AMP | APsL |
|-----------|-------------|--------|--------|----|------|------|-----|-----|------|
| RetinaNet | ResNet-50   | 37.74M | 239.32 | 35.6 | 55.2 | 38   | 19.6| 39.4 | 47   |
|           | +SE block   | 40.23M | 239.43 | 36  | 55.8 | 38.6 | 19.5| 39.9 | 47.2 |
|           | +ECA block  | 37.74M | 239.43 | 36.6| 56.5 | 39.2 | 20.8| 40.6 | 48.8 |
|           | +DCA(Ours)  | 37.74M | 239.43 | 36.8| 56.8 | 39.4 | 21.1| 40.7 | 48.5 |
| Mask R-CNN| ResNet-50   | 44.18M | 275.58 | 37.4| 58.9 | 40.7 | 22.1| 40.9 | 48.3 |
|           | +SE block   | 46.67M | 275.69 | 37.8| 59.9 | 41.2 | 22.7| 42   | 48.6 |
|           | +ECA block  | 44.18M | 275.69 | 38.3| 60.5 | 42   | 22.7| 42   | 49.7 |
|           | +DCA(Ours)  | 44.18M | 275.69 | 38.3| 60.4 | 41.4 | 23.3| 42.5 | 49.2 |

Table 4. Instance segmentation results of different methods using Mask R-CNN on COCO val2017.

| Methods | AP | AP50 | AP75 | APS | AMP | APsL |
|---------|----|------|------|-----|-----|------|
| ResNet-50| 34.1| 55.6 | 36   | 18.3| 37.4| 46.5 |
| +SE block | 34.7| 56.5 | 37   | 18.9| 38.5| 47   |
| +ECA block | 35.1| 57.1 | 37.1 | 19  | 38.9| 47.4 |
| +DCA(Ours)  | 35.1| 57   | 37.3 | 18.8| 38.9| 47.6 |

4.6. Image Segmentation on MS-COCO
Finally, we use Mask R-CNN to verify the performance of DCA-Net on instance segmentation. As shown in Table 4, the AP of DCA-Net is the same as ECA-Net, specifically, it is 1% higher than the original ResNet-50. Compared with ECA-Net, the AP75 and APsL of DCA-Net are both improved by 0.2%.

All the above results prove that our DCA-Net has versatility for various tasks of computer vision.

5. Conclusion
In this paper, we propose a novel lightweight attention network called DCA-Net. DCA-Net consists of a pre-attention module and a post-attention module. By introducing a pre-attention module, DCA-Net solves the problem that ECA-Net cannot obtain the global channel interaction relationships and
achieves better performance in image classification, object detection, and instance segmentation tasks. At the same time, despite the addition of two additional modules, the parameter amount of DCA-Net is still less than ECA-Net. In the future, we will simplify the structure of the model further and consider combining channel attention and spatial attention together.

Reference

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. “Imagenet Classification with Deep Convolutional Neural Networks.” In Advances in Neural Information Processing Systems, 1097–1105

[2] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. “Going Deeper with Convolutions.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–9

[3] Karen Simonyan, and Andrew Zisserman. 2014. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” arXiv Preprint arXiv:1409.1556

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. “Deep Residual Learning for Image Recognition.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778

[5] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. “Densely Connected Convolutional Networks.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4700–4708

[6] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. 2017. “Factorized Bilinear Models for Image Recognition.” In Proceedings of the IEEE International Conference on Computer Vision, 2079–2087

[7] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. “Non-Local Neural Networks.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7794–7803

[8] Jie Hu, Li Shen, and Gang Sun. 2018. “Squeeze-and-Excitation Networks.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7132–7141

[9] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. “Cbam: Convolutional Block Attention Module.” In Proceedings of the European Conference on Computer Vision, 3–19

[10] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. 2018. “A^2-Nets: Double Attention Networks.” In Advances in Neural Information Processing Systems, 352–361

[11] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo and Qinghua Hu. 2020. “ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks.” In The IEEE Conference on Computer Vision and Pattern Recognition

[12] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. “Neural Machine Translation by Jointly Learning to Align and Translate,” arXiv Preprint arXiv:1409.0473

[13] Alex Krizhevsky, Geoffrey Hinton, and others. 2009. “Learning Multiple Layers of Features from Tiny Images.” http://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. “Microsoft Coco: Common Objects in Context.” In European Conference on Computer Vision, 740–755. Springer.

[15] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. 2019. “Dual Attention Network for Scene Segmentation.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3146–3154

[16] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. “Aggregated Residual Transformations for Deep Neural Networks.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1492–1500
[17] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, et al. 2019. “MMDetection: Open Mmlab Detection Toolbox and Benchmark.” arXiv Preprint arXiv:1906.07155.

[18] Terrance DeVries, and Graham W Taylor. 2017. “Improved Regularization of Convolutional Neural Networks with Cutout.” arXiv Preprint arXiv:1708.04552.

[19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. “Feature Pyramid Networks for Object Detection.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2980-2988.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. “Mask R-Cnn.” In Proceedings of the IEEE International Conference on Computer Vision, 2961–2969.

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. “Focal Loss for Dense Object Detection.” In Proceedings of the IEEE International Conference on Computer Vision, 2117-2125.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. “Faster R-Cnn: Towards Real-Time Object Detection with Region Proposal Networks.” In Advances in Neural Information Processing Systems, 91–99.