

# EFFICIENT MULTI-SCALE ATTENTION MODULE WITH CROSS-SPATIAL LEARNING

*Daliang Ouyang, Su He, Guozhong Zhang, Mingzhu Luo, Huaiyong Guo, Jian Zhan, Zhijie Huang*

AEROSPACE SCIENCE & INDUSTRY SHENZHEN (GROUP) CO., LTD.,

Shenzhen, China

Email: [{ouyangdl, hesu, zhanggz, luomz, guohy, zhanj, huangzj}@casic.com.cn](mailto:{ouyangdl, hesu, zhanggz, luomz, guohy, zhanj, huangzj}@casic.com.cn)

## ABSTRACT

Remarkable effectiveness of the channel or spatial attention mechanisms for producing more discernible feature representation are illustrated in various computer vision tasks. However, modeling the cross-channel relationships with channel dimensionality reduction may bring side effect in extracting deep visual representations. In this paper, a novel efficient multi-scale attention (EMA) module is proposed. Focusing on retaining the information on per channel and decreasing the computational overhead, EMA groups the channel dimensions into multiple sub-features and makes the spatial semantic features well-distributed inside each feature group. Specifically, apart from encoding the global information to re-calibrate the channel-wise weight in each parallel branch, the output features of the two parallel branches are further aggregated by a cross-dimension interaction method. The extensive experiments on common-used benchmarks, such as CIFAR100 for image classification, and object detection on MS COCO and VisDrone2019 datasets, are conducted which indicate that EMA outperforms several recent attention mechanisms significantly without changing networks depth.

**Index Terms**— Attention mechanism, cross-dimension interaction, image classification, object detection

## 1. INTRODUCTION

Following the evolution of deep Convolutional Neural Networks (CNNs), more and more notable network topologies are employed in the fields of image classification and object detection tasks for obtaining superior performance. However, it may bring a number of additional parameters and calculations, when we extend CNNs to across multiple convolutional layers for enhancing the learnt feature representation [1], [2]. The attention mechanism, due to the flexible structure characteristics that can be easily plugged into backbone architecture of CNNs, attracts much interest in the research communities of computer vision.

As the representative channel attention, Squeeze-and-excitation (SE) explicitly modeled the interdependencies between inner-channels with the goal of strengthening the representational power of CNNs [3]. In Convolutional block

attention module (CBAM) [4], it established the model with the semantic interdependencies between spatial and channel dimensions and exploited both spatial and channel attentions in an extremely efficient way. For overcoming the computational cost limitations, a long-standing and effective way that using feature grouping to divide features into multi-group on different resources is provided [5]. Based on this, Spatial group-wise enhance (SGE) attention [6] grouped the channel dimensions into multiple sub-features and improved the spatial distribution of different semantic sub-features representations. The convolution with channel dimensionality reduction is also an effective way to control model complexity [7]. By modifying from SE, Coordinate attention (CA) embedded the positional information into the channel attention and selected an appropriate reduction ratio of channel dimensionality to overcome the paradox of performance and complexity trade-off [8]. Although choosing the appropriate channel reduction ratios yields better performance, it may bring side effect in extracting deep visual representations, which is explored the efficiency in Efficient channel attention (ECA) [9].

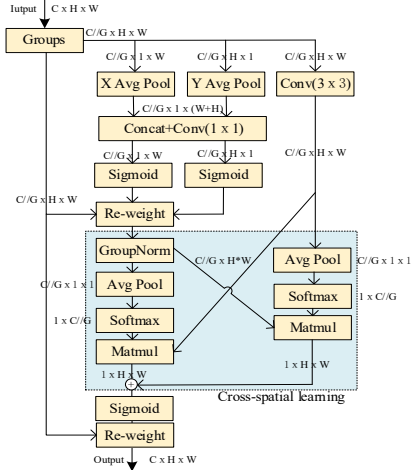
Taking the inspiration from the aforementioned attention mechanisms, we, based on the feature grouping, revise the sequential processing method of CA and propose a novel efficient multi-scale attention (EMA) over the cross-spatial learning method. The experimental results indicate that EMA imposes only a slight increase in computational burden and improves the performance for CNNs with different architectures and depths. Our main contributions are concluded as follows:

- We propose a novel cross-spatial learning method and design a multi-scale parallel subnetworks for establishing both short and long-range dependencies.
- We consider a generic method that reshapes the partly channel dimensions into the batch dimension to avoid some form of dimensionality reduction via a universal convolution.
- Apart from building the local cross-channel interaction in each parallel subnetwork without channel dimensionality reduction, we also fuse the output feature maps of the two parallel subnetworks by a cross-spatial learning method.

## 2. RELATED WORK

Large layers depth plays an important role in increasing the representational ability of CNNs, but it inevitably leads to more sequential processing and higher latency [10], [11]. Different from the large depth attentions described as a linear sequence of stacking more layers, Triplet attention [12] proposed a triplet parallel branches structure for capturing cross-dimension interaction against the parallel branches. With the parallel structures, Shuffle attention (SA) [13] grouped channel dimensions into multiple sub-features, which can be efficiently parallelized across multiple processors. However, only part of channels will be taken into account to exploit the inter-relationship of channels and construct informative features by fusing both spatial and channel-wise information.

Different to the above attention methods, where the learnt attention weights are aggregated by a simple averaging method as the final output of the parallel subnetworks, we fuse the output feature maps of the parallel subnetworks by a cross-spatial learning method that uses the matrix dot-product operations aiming at capturing pixel-level pairwise relationship and highlighting global context for all pixels [14], [15].



**Fig.1.** It shows the structures of our proposed EMA module. Here, “G” means the divided groups, “X Avg Pool” represents the 1D horizontal global pooling and “Y Avg Pool” indicates the 1D vertical global pooling, respectively.

## 3. EFFICIENT MULTI-SCALE ATTENTION (EMA)

The overall structure of EMA module is shown in Figure. 1. In this section, we will discuss how the EMA learns effective channel descriptions without channel dimensionality reduction and produces a better pixel-level attention for high-level feature maps. Specifically, we only pick out the shared component of the 1x1 convolution from CA and we name it as 1x1 branch in EMA.

For any given input feature map  $X \in \mathbb{R}^{C \times H \times W}$ , EMA will divide  $X$  into  $G$  sub-features across the channel

dimensions direction, where  $C$  means the numbers of the input channels,  $H$  and  $W$  indicate the spatial dimensions of the input features respectively. Given the truth that there are no batch coefficients in the dimension of the convolution function for the normal convolution, the number of convolution kernels are independent of the batch coefficient of the forward operational inputs. To make it more concrete, the input dimensions of the normal 2D convolution kernel in Pytorch is  $[oup, inp, k, k]$ , where  $oup$  means the out planes of the inputs,  $inp$  indicates the input planes of the input features and  $k$  denotes the kernel size respectively. Obviously, it is not involved the batch dimensions with convolution operations. Accordingly, we reshape and permute  $G$  groups into the batch dimensions, and redefine the input tensors with shape of  $C // G \times H \times W$ . The groups-style can be redefined as  $X = [X_0, X_1, \dots, X_{G-1}]$ ,  $X_i \in \mathbb{R}^{C // G \times H \times W}$ . Without losing generality, we let  $G \ll C$ .

In EMA, it conducts that three parallel routes are exploited to extract attention weight descriptors of the grouped feature maps. Two of the parallel routes is in 1x1 branch and the third one route is that the 3x3 branch. To be more specific, there are two 1D global average pooling operations employed to encode the channel attention along two spatial directions respectively in 1x1 branch, and only one single 3x3 kernel is stacked in 3x3 branch for capturing multi-scale feature representation. One the one hand, EMA will decompose the original input tensors into two parallel 1D feature encoding vectors for modeling the remote long-range spatial dependencies. Firstly, one of the parallel processing routes is directly from a 1D global average-pooling along the horizontal dimension direction and hence can be viewed as a collection of positional information along the vertical dimension direction [9]. Consequently, the 1D global average-pooling for encoding the global information along the horizontal dimension direction in  $C$  at  $H$  can be denoted by

$$z_c^H(H) = \frac{1}{W} \sum_{0 \leq i \leq W} x_c(H, i) \quad (1)$$

where  $x_c$  indicates the input features at  $c$ -th channel. With such encoding processes, EMA can capture the long-range dependencies at the horizontal direction and preserve precise positional information at the vertical direction. Similarly, the other one parallel route is directly from 1D global average-pooling along the horizontal dimension direction and hence can be viewed as a collection of positional information along the vertical dimension direction. Then, the route utilizes the 1D global average-pooling along the vertical dimension direction to capture long-range interactions spatially and preserve the precise positional information along the horizontal dimension direction. The pooling output in  $C$  at  $W$  can be formulated as

$$z_c^W(W) = \frac{1}{H} \sum_{0 \leq j \leq H} x_c(j, W) \quad (2)$$

where  $x_c$  indicates the input features at  $c$ -th channel. In the following, the input features can encode the global feature information and assist the model in capturing global information along two spatial directions respectively. We concatenate the two encoded features against the images height direction and make it share the same 1x1 convolution in 1x1 branch. After factorize the outputs of 1x1 convolution into two 1D vectors, two non-linear Sigmoid functions are employed to fit the 2D binomial distribution upon linear convolutions. And then, we can aggregate the two channel attention maps inside each group via a simple multiplication which adaptively recalibrate the channel-wise relationship for achieving different cross-channel interactive features between the two parallel routes in 1x1 branch. Since the preserved precise positional information along the different spatial direction is complementary, recalibrating the raw input features makes EMA learn low-level detailed feature representations. On the other hand, the raw input features will also pass through the 3x3 branch via a 3x3 convolution to enlarge the feature space.

Benefiting from the capability of building interdependencies among channels and spatial locations, there have been extensively studied and broadly used in a variety of computer vision tasks recently [16], [17]. Inspired by this, we provide a cross-spatial information aggregation method at different spatial dimension direction for richer feature aggregation. Note that here, we still have introduced two branches where one is the output of 1x1 branch and the other is the output of the 3x3 branch. Correspondingly, we utilize the 2D global average pooling to encode global spatial information in 1x1 branch and the outputs of the least branch will be transformed to the correspond dimension shape directly before the joint activation mechanism of channel features, i.e.,  $\mathbb{R}_1^{1 \times C // G} \times \mathbb{R}_3^{C // G \times HW}$  [18]. The 2D global pooling operation is formulated as

$$z_c = \frac{1}{H \times W} \sum_j^H \sum_i^W x_c(i, j) \quad (3)$$

which is designed for encoding the global information and modeling the long-range dependencies. For efficient computation, the natural non-linear functions Softmax for 2D Gaussian maps is employed at the outputs of 2D global average pooling to fit the upon linear transformations. By multiplying the outputs of above parallel processing with matrix dot-product operations, we derived our first spatial attention map, which collects different scale spatial information in the same processing stage. Moreover, we similarly utilize the 2D global average pooling to encode global spatial information in 3x3 branch and the 1x1 branch will be transformed to the correspond dimension shape directly before the joint activation mechanism of channel features, i.e.,  $\mathbb{R}_3^{1 \times C // G} \times \mathbb{R}_1^{C // G \times HW}$ . After that, the second spatial attention map which preserves the entire precise spatial positional information is derived. Finally, the output feature maps within each group are calculated as the aggregation of the two generated spatial attention weight

values followed by a Sigmoid function. The final output of EMA is the same size of  $X$ , which is efficient yet effective to stack into modern architectures. Considering the cross-spatial information aggregation method, both the long-range dependencies will be modeled and the precise positional information are embedded into EMA emphasizing the representations of interests [19].

**Table 1:** Comparison of various attention methods on CIFAR100. Note that all of the attention modules are embedded into the bottleneck blocks of ResNet50/101.

Method	Backbone	#.Param.	FLOPs	Top-1 (%)	Top-5 (%)
Baseline	ResNet50	23.71M	1.30G	77.26	93.63
+ CBAM [20]		26.24M	1.31G	80.56	95.34
+ SA		23.71M	1.31G	79.92	95.00
+ ECA		23.71M	1.31G	79.68	95.05
+ NAM [20]		23.71M	1.31G	80.62	95.28
+ CA		25.57M	1.36G	80.17	94.94
+ EMA (ours)		<b>23.85M</b>	<b>1.32G</b>	<b>80.69</b>	<b>95.59</b>
Baseline	ResNet101	42.70M	2.53G	77.78	94.39
+ CA		46.22M	2.54G	80.01	94.78
+ EMA (ours)		<b>42.96M</b>	<b>2.53G</b>	<b>80.86</b>	<b>95.75</b>

## 4. EXPERIMENTS

In this section, we provide the details for experiments and the results demonstrate the performance and efficiency of our proposed EMA.

### 4.1. Image Classification on CIFAR100

We investigate our proposed EMA on CIFAR100 datasets, whose sets include the images with 32x32 pixels and consist of images drawn from 100 classes. We exploit stochastic gradient descent (SGD) with momentum of 0.9 and the weight decay of 4e-5. The batch size is 128 by default. Our networks of all comparing approaches are trained for 200 epochs and are performed with exactly the same training configuration settings in NAM [20] to make fair comparisons. We evaluate the performance of our network using the standard Top-1 and Top-5 accuracy metrics.

As shown in Table 1, a comparison of several other attention mechanisms over the baseline of ResNet50/101 [21] shows that integrating with EMA gains a very comparative performance with relatively small model complexity (i.e., network parameters and floating-point operations per second (FLOPs)). Comparing with the standard baseline of ResNet50, EMA achieves 3.43% gains in terms of Top-1 accuracy and 1.96% advantages over the Top-5 accuracy. With almost the same computational complexity, the Top-1 accuracy can be improved by 0.52% by our proposed EMA as compared to the CA. In addition, using ResNet101 as the backbone model, we compare EMA with CA. Obviously, our EMA outperforms CA by a large margin with less parameters (42.96M v.s. 46.22M) and lower computational cost. It is worth noting that the gain in term of Top-1 average accuracy of CA is slightly dropped

from 80.17% for the ResNet50 to 80.01% for the ResNet101 as the network architecture becoming deeper.

**Table 2:** Object detection results of different attention methods on COCO and VisDrone val datasets.

Model	Datasets	#.Param.	FLOPs	mAP (0.5)	mAP (0.5:0.95)
Yolov5s	COCO	7.23M	16.5M	56.0	37.2
+ CBAM		7.27M	16.6M	57.1	37.7
+ SA		7.23M	16.5M	56.8	37.4
+ ECA		7.23M	16.5M	57.1	37.6
+ CA		7.26M	16.50M	57.5	38.1
+ EMA (ours)		<b>7.24M</b>	<b>16.53M</b>	<b>57.8</b>	<b>38.4</b>
Yolov5x	VisDrone	90.96M	314.2M	49.29	30.0
+ CBAM		91.31M	315.1M	49.40	30.1
+ CA		91.28M	315.2M	49.30	30.1
+ EMA (ours)		<b>91.18M</b>	<b>315.0M</b>	<b>49.70</b>	<b>30.4</b>

## 4.2. Object Detection on MS COCO and VisDrone

For the training, apart from the fixed size of input demanded by the original YOLOv5s [22], we select the configuration settings such as resizing the images to uniform dimensions of  $640 \times 640$ . The epochs and the batchsize are set as 300 and 48 respectively. The model performance is evaluated by using the metrics, such as model size, parameters, FLOPs and the accuracy of mAP (0.5) (threshold of 0.5) and mAP (0.5:0.95) (threshold from 0.5 to 0.95).

As shown in Table 2, we can see that the integration of either EMA or the other attention modules into Yolov5s backbone both improve the performance by a clear margin. Compared to CBAM, SA and ECA, EMA exceeds the baseline of YOLOv5s by much performance gaining and performs slightly better than the CA in terms of mAP (0.5). For example, the model size of EMA is 7.24M, which is only 0.01M lightly larger than the baseline of YOLOv5s, ECA and SA. Although the FLOPs of EMA are 16.53M, which are only 0.03M larger than the baseline of YOLOv5s, EMA achieves the mAP (0.5) of 57.8% and mAP (0.5:0.95) of 38.4% on all 80 classes.

Considering our proposed EMA on the dense object detection of the multi-scale feature fusion, we add a detection head for the tiny objects based on the original YOLOv5x [23] and integrate EMA into prediction branch to achieve the purpose of exploring the prediction potential. During experiments, we set the size of the input image to  $640 \times 640$  and we use part of pre-trained model from yolov5x for saving a lot of training time. All the attention models on VisDrone2019 trainset are trained for 300 epochs and the batch size is set as 5. The experiments are performed with exactly the same training configuration settings. We use the YOLOv5x as our backbone CNN for the object detection, where the CA, CBAM and EMA attentions are integrated into the detector respectively. It is noteworthy that CBAM boosts the performance of YOLOv5x by 0.11% and is higher than that of CA at the cost of more parameters and computations. For the CA, it almost obtains the same

performance as the baseline and surpasses the YOLOv5x by 0.01% in terms of the mAP (0.5), while CA achieves higher parameters and computations than EMA (91.28M v.s. 91.18M and 315.2M v.s. 315.0M). Specifically, EMA adds 0.22M more parameters than baseline method, which have an improvement of 0.41% over YOLOv5x on mAP (0.5) and 0.4% on mAP (0.5:0.95) with the slightly higher parameters. These results demonstrate that EMA is an efficient module for object detection task, and further proves the effectiveness of the EMA method in this paper.

## 5. ABLATION STUDY

We choose ResNet50 as baseline networks and validate the importance of cross-spatial learning method by conducting ablation experiments to observe the impact of different hyperparameters in EMA, such as EMA\_no (no cross-spatial learning), EMA\_16 (group size is 16) and EMA\_32 (group size is 32). Comparing with EMA\_32, a relatively high FLOPs and network parameters will be resulted by setting group size as 16. This is mainly due to reshape the channel dimensions into the batch dimensions that decreases the model parameters, EMA is able to call upon to distribute the model over multiple channels on more batch dimensions and process them. Moreover, we also conduct the ablation study by conducting cross-spatial learning method and the other turns off. From the view of Table 3, EMA\_32 with the cross-spatial learning method outperforms EMA\_no scheme. For the similarly FLOPs and network parameters, the Top-1 and Top-5 rates of EMA\_32 are much higher, at 80.69% and 95.59%, respectively.

**Table 3:** Ablation on relative training configuration settings.

Method	Datasets	#.Param.	FLOPs	Top-1 (%)	Top-5 (%)
+ EMA_no	CIFAR 100	23.84M	1.32G	78.24	94.89
+ EMA_16		24.44M	1.34G	80.35	95.44
+ EMA_32		<b>23.84M</b>	<b>1.32G</b>	<b>80.69</b>	<b>95.59</b>

## 6. CONCLUSION

In this paper, we systematically investigate the properties of attention mechanisms and present new insight into how the CNNs can enjoy both good generalization and computation budgets, by using a generic method that avoids some form of dimensionality reduction and contribute to capture long-range feature interdependencies via a universal convolution. Due to the flexible and light-weighted characteristics, our proposed EMA can be easily exploited into different computer vision tasks. Note that this paper currently focuses on image classification, and object detection for model development. However, we believe EMA is more applicable to broader applications like semantic segmentation and can be stacked into other deep CNNs for significantly enhancing the feature representation ability. We will leave them for future work.

## 7. REFERENCES

- [1] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5659–5667.
- [2] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.
- [3] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
- [4] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
- [5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [6] Xiang Li, Xiaolin Hu, and Jian Yang, "Spatial group-wise enhance: Improving semantic feature learning in convolutional networks," *arXiv preprint arXiv:1905.09646*, 2019.
- [7] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 116–131.
- [8] Qibin Hou, Daquan Zhou, and Jiashi Feng, "Coordinate attention for efficient mobile network design," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 13713–13722.
- [9] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11534–11542.
- [10] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.
- [11] Xiang Li, Wenhui Wang, Xiaolin Hu, and Jian Yang, "Selective kernel networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 510–519.
- [12] Diganta Misra, Trikey Nalamada, Ajay Uppili Arasanipalai, and Qibin Hou, "Rotate to attend: Convolutional triplet attention module," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3139–3148.
- [13] Qing-Long Zhang and Yu-Bin Yang, "Sa-net: Shuffle attention for deep convolutional neural networks," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 2235–2239.
- [14] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, "Dual attention network for scene segmentation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3146–3154.
- [15] Luchen Liu, Sheng Guo, Weilin Huang, and Matthew R Scott, "Decoupling category-wise independence and relevance with self-attention for multi-label image classification," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 1682–1686.
- [16] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng, "A<sup>2</sup>-nets: Double attention networks," *Advances in neural information processing systems*, vol. 31, 2018.
- [17] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.
- [18] Huajun Liu, Fuqiang Liu, Xinyi Fan, and Dong Huang, "Polarized self-attention: Towards high-quality pixelwise regression," *arXiv preprint arXiv:2107.00782*, 2021.
- [19] Guangxiang Zhao, Xu Sun, Jingjing Xu, Zhiyuan Zhang, and Liangchen Luo, "Muse: Parallel multi-scale attention for sequence to sequence learning," *arXiv preprint arXiv:1911.09483*, 2019.
- [20] Yichao Liu, Zongru Shao, Yueyang Teng, and Nico Hoffmann, "Nam: Normalization-based attention module," *arXiv preprint arXiv:2111.12419*, 2021.
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [22] Glenn Jocher, Alex Stoken, Ayush Chaurasia, Jirka Borovec, Yonghye Kwon, K Michael, C Liu, J Fang, V Abhiram, SP Skalski, et al., "ultralytics/yolov5: v6.0—yolov5n 'nano' models, roboflow integration, tensorflow export, opencv dnn support," *Zenodo Tech. Rep.*, 2021.
- [23] Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao, "Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 2778–2788.