IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 7, 2021, accepted March 17, 2021, date of publication March 24, 2021, date of current version March 31, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068488

Att-FPA: Boosting Feature Perceive for Object Detection

JINGWEI LIU[1,2], YI GU[3], SHUMIN HAN[3], ZHIBIN ZHANG[1,2], (Member, IEEE), JIAFENG GUO[1,2], (Member, IEEE), AND XUEQI CHENG[1,2], (Senior Member, IEEE)

[1]CAS Key Laboratory of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Baidu Online Network Technology (Beijing) Company Ltd., Beijing 101399, China

Corresponding author: Shumin Han (hanshumin@baidu.com)

ABSTRACT The deep convolutional networks have a great success in vision classification tasks. For object detection, tasks are divided into two subtasks: localization and classification. The detectors scan the whole image to generate object proposals relying on the predefined anchors or points, then classify and fine trim the proposals. The localization task plays an important role in object detection. The foreground objects and background can be easily confused under complex scenes in the existing approaches. Thus, the localization task results in a bad influence on the performance of classification problems. In order to enhance the object localization perceived, a novel method called Attentional Feature Perceive and Augmentation (Att-FPA) is constructed, which devotes setting up a feature dual perceive for localization and classification. Firstly, an attention mask branch (with an attention mask) is imported to check out if the detecting area is the background or containing one or more objects. By maintaining the feature representation of background, the attention mask is employed to strengthen the feature representation of foreground objects, which significantly enhances the localization perceived ability of objects. Moreover, Att-FPA supplies a novel way on feature re-extraction and augmentation, which further promotes the performance significantly. In the task of Ms COCO object detection, Att-FPA achieves an outstanding promotion(3.0+ mAP) than baseline (Faster-Rcnn) method. It establishes a new efficient method of feature representation, which outperforms the state-of-the-art models in object detection.

INDEX TERMS Deep learning, computer vision, object detection, attention, feature representation.

I. INTRODUCTION

Convolutional neural networks(CNNs) [1] have met with great success in image classification due to its strong ability of image feature representation. It is natural that CNNs are leading to work out the task of object detection [2]–[7]. The tasks of object detection are divided into two sub-tasks: classification and localization. Different detection architectures are designed based on CNNs to improve the performance of object detection tasks have been proposed in recent years, mainly addressed in two ways: two-stage [8]–[10] and one-stage framework [11]–[14]. In two-stage framework, the region proposal network(RPN) [10] is added and the coarse localization is detected. After that, the classification and fine localization are finished in the second stage [10].

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

The commonality of the two sub-tasks is that they are sharing the same Backbone and Necks [15]. Feature extraction is mainly completed in Feature Pyramid Network (FPN) [8], which plays a very important role in detection tasks. FPN is a simple and effective top-down structure to solve the multi-scale problem. However, as mentioned in Aug-FPN [16], there are several disadvantages due to the scale-decrease model, such as the detailed gaps of semantic and geometric between different feature levels. Because of the ability of feature expression becomes weaker in high level [16], the features with small response are ignore during feature fusion operation. [17].

Recently, many approaches have been proposed to eliminate the flaws and achieve significant improvements on scale-permuted, multi-scale feature fusion and scale-equalizing, such as Spinenet [17], Aug-FPN [16], Efficientnet [19], ASFF [20] and SEPC [21]. However, all methods operate on

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

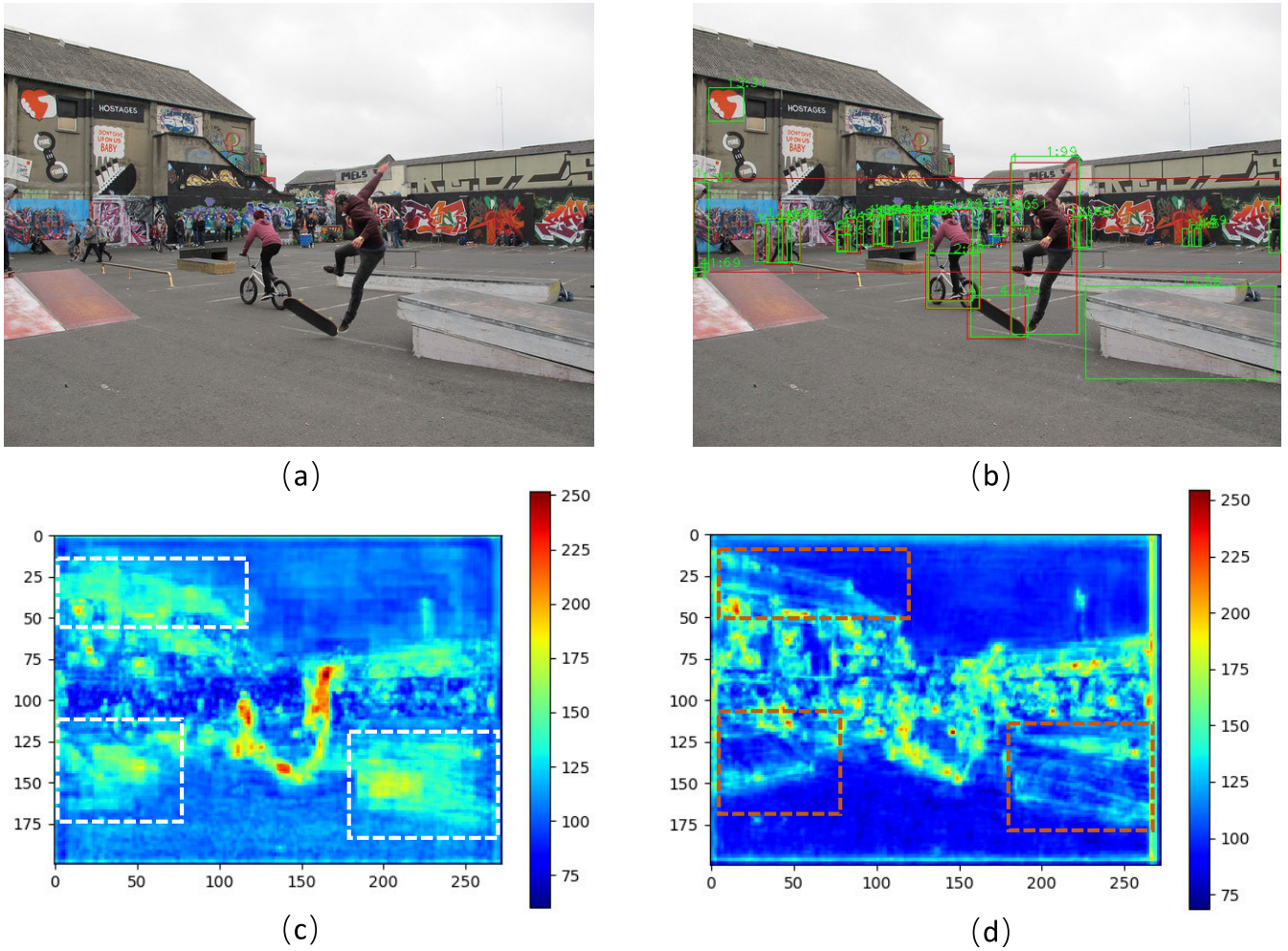


FIGURE 1. Illustration of indistinguishability between features of foreground objects and background: (a) Input image comes from Ms COCO datasets *val*2017 which has lots of objects and complex background; (b) The result of detection on Faster-Rcnn-Resnet101 [10], the red bounding boxes are ground truth marking manually. The green bounding boxes are the output of detectors (the score of NMS [18] is 0.05), the classify and score are shown in the head of bounding box separated by colon. (c) The feature map of (b); (d) The feature map of the proposed Att-FPA based on Faster-Rcnn-Resnet101. In (c), the areas on the rectangle of white dashes belong to the background, but they have high responses and cause the error detection. As (d) shows, the same areas(red rectangles) have a low response. Compared with (c), (d) has a more clear background and meaningful presentation (refection on shape and texture) on objects and background.

a whole feature map, which ignore the discrepancy between the foreground and background objects. If the background features are regarded as the same as the foreground ones, the operation would obtain an higher false positive(FP) and the representation of object feature will be decreased.

Three challenges are described as follows to overcome for improving the object detection on feature extraction:

A. THE DIFFICULTIES TO DISTINGUISH THE FEATURES AT THE FOREGROUND OBJECTS OR BACKGROUND

In Fig. 1, as the input image shows, person, bicycle, skateboard are the objects we need to detect in the detection task on Ms COCO datasets [34]. The asphalted road, sky, houses and others are regarded as the background in this image. Simultaneously, the advertising slogan, cement flag pavement, sundries are also background in the detection task. Here, we call them pseudo objects. Unfortunately, these pseudo objects are always detected as the general foreground objects in the previous methods. As (b) shows, the advertising slogan (red heart in hand) and cement flag pavement are detected with the confidence of 0.3 and 0.56, these pseudo objects bring challenges which confuse us. The phenomenon can be illustrated in (c), as the feature map shows, these pseudo objects (marked by the white bounding boxes) have an approximately responsiveness as the foreground objects. How to suppress the responsiveness of pseudo objects becomes a challenge.

B. THE DIFFERENCE BETWEEN SUB-TASKS OF LOCALIZATION AND CLASSIFICATION

In general object detection architecture, the tasks of localization and classification are always trained based on the same backbone and neck. As mentioned in [27], the sub-task of classification relies on the features which have a good presentation on the texture and color of images. Different

from classification, the sub-task of localization infers encode spatial position relies sufficiently on such cues. It is obvious that the semantic information in the feature map is more useful for classification, and the geometrical information is more useful for localization. In Fig. 1, (c) shows that, the semantic and geometrical details of objects (person, bicycle) are inaccuracy from a visual standpoint. In object detection, the features which are extracted from one shared backbone and necks are hardly obtain an optimal results in both the problems of classification and localization. How to keep a trade-off between these two sub-tasks is a new challenge in the object detection.

C. WEAK REPRESENTATION IN RoIs

In RPN, the predefined anchors and points are classified as positive or negative samples. The rectangular proposals with scores are generated by predefined anchors or points. The features of proposals are produced by regional of interests (RoI) Pooling. However, the RoI is cropped from the layers of FPN, which results in weak expression. For small and tiny objects, the features representation ability limited due to the scale decrease in FPN. The features of large objects are lost due to the max pooling operation. The weak representation is one of the factors on reducing the performance.

Our proposed method aims at obtaining a reasonable feature extraction approach to distinguish the foreground objects and background in various object detection tasks. In this paper, an Attentional Feature Perceived and Augmentation (Att-FPA) is designed for benefiting both the classification and localization problems. The features of the proposed approach are shown as follows:

- During the feature fusion procedure, the foreground objects and background are treated as the same in each feature map, which lack of meaningful discrimination. The general method can result in an undesired high concentrating on background, which easily brings a higher FP. Therefore, it is expected that the designed CNNs model can pay more attention on feature representation of foreground objects. For the first challenge above, an attention mask is introduced to strength the ability of discrimination between foreground objects and background. The attention mask is constrained by the ground truth mask. And the ground truth boxes of manual annotation are used for generating the ground truth mask. In Fig. 1, (d) shows the feature map generated by the attention mask (the blue color denotes low responsiveness and the red denotes high responsiveness). The feature map of sky and asphalted road are more clear without interferential responsiveness. The feature map of advertising slogan and cement flag pavement has low but meaningful responsiveness (semantic and geometric details).
- In the corporation with the attention mask, we import another feature pyramid. The second challenge is resolved by dual sub-branch feature pyramid. One feature pyramid branch with attention mask is designed for the localization task. Another feature pyramid branch with better semantic information is designed for classification sub-task. In Fig. 1, compared with (c), (d) shows that the contour of objects (person, bicycle, skateboard, ridge pole and so on) are more reasonable.
- Regarding the third challenge, the features of proposals are re-extracted from the backbone and neck based on the proposal image (cropped from the original image by proposal coordinates). In general, the backbone (CNNs) has a good representation on semantic information. Because of this, the features of proposals are extracted again by the shared backbone and neck. The large and small objects maintain their original information. The original and re-extracted features are combined for further classifying and fine trimming of localization.

The **main contributions** of this paper are:

1) An attention mask based on the ground truth is imported, the attention mask captures the geometrical information and provide meaningful differentiable features between the foreground objects and background.
2) A novel dual feature pyramid network is designed to serve for classification and localization.
3) The information is extracted from backbone and neck for RoI feature augmentation. With the proposed approach, the mean Average Precision (mAP) on Ms COCO increased by 3.1% without any data augmentation and tricks.

II. RELATED WORK

A. FEATURE EXTRACTION AND FUSION ON OBJECT DETECTION

Object detection is a basic task in computer visions. Two-stage and one-stage are the two kinds of general frameworks for designing the object detectors. In both two frameworks, the feature of objects and background are extracted from the backbone and neck. The representation of the features plays an important role in object detection. However, different levels of feature have a gap on semantic information and geometric details. To eliminate the gap, in PAFPN [22] and ASFF [20], the high-level and low-level feature map are combined by employing top-down and bottom-up information flow paths. In NAS-FPN [23], the NAS algorithm is leveraged to fuse features across scales, which also consists of a combination of top-down and bottom-up connections. In EfficientDet [19], the BiFPN module is proposed for feature fusion. In SEPC [21], a scale-equalizing pyramid convolution is designed based on Scale-space theory to relax the discrepancy between the feature pyramid and the Gaussian pyramid. The intrinsic flaw around feature fusion is systematically analyzed in AugFPN [16]. An adaptive channel fusion, Residual Feature Augmentation, and Soft RoI Selection are proposed for feature extraction and fusion. In [24], CBNet engages the higher-level features flow into the parallel stage of the succeeding backbone. Besides, the spatial sensitivity

of classification and localization on the output feature maps are explored in TSD [25].

B. ATTENTION MECHANISMS

Attention mechanisms are widely used in Natural language processing (NLP) [26]. In recent years, attention modules are employed in various vision tasks, such as image classification, image captioning, object detection, scene segmentation, and human pose estimation. An attention mechanism is proposed in [28], which learns to softly weight the multi-scale features at each pixel location. Then, [29] verified that self-attention can be an effective stand-alone layer. The proposed pure self-attention model gains an outstanding performance and fewer FLOPS on Ms COCO object detection. [30] proposed an augmented convolutional operator with this self-attention mechanism. In [31], a novel Dual Attention Network (DANET) with self-attention mechanism is proposed to strength the discriminant ability of feature representations for scene segmentation.

In this paper, different from the above attention mechanism which is hard to train and optimize, the proposed attention mechanisms are guided by the ground truth. It is much more convenient on object perceiving and maintaining the background information. Besides, the proposed attention mechanisms bring less time consumption on calculation.

III. PROPOSED APPROACH

A. OVERVIEW

Based on the two-stage framework, the proposed approach with attention based feature perceiving and augmentation is proposed in this paper, which focuses on building a strong ability of feature representation for classification and localization sub-task. The proposed Att-FPA distinguishes the foreground objects from background effectively, and obtains a strong feature representation. The pipeline of our method is illustrated in Fig. 2. In the FPN model, feature map layers L in branch A and feature map layers C branch B are generated by backbone (resnet [37], vgg [38] and so on). The feature map layers C is masked by the attention mask, and jointed with L as the output of FPN, which is perceived and distinguishable.

Multi-scale layers of feature map are the input of RPN, the predefined anchors and points are tiled on them and plenty of proposal bounding boxes are generated. The input image is re-cropped as RoI images by the above proposal bounding box under specific filter conditions F. The RoI images are resized into a fixed small size and fed into the backbone and necks to generate the augmented features R. In Fig. 2, the dash line in black shows the classic procession and the blue lines in solid denote the process of feature re-extraction. In the second stage, the origin RoI features O are produced by RoI pooling layer. The features R are strengthened by O and used for the RCNN.

B. DUAL FEATURE PYRAMID

The image is fed into backbone and the feature maps pyramid $\{C_2, C_3, C_4, C_5\}$ with different channels are the output. In this paper, we introduce the dual feature pyramid branch A and branch B with different convolution filters. Here, branch A is denoted as $\{A_M_2, A_M_3, A_M_4, A_M_5, A_M_6\}$, and branch B is denoted as $\{B_M_2, B_M_3, B_M_4, B_M_5, B_M_6\}$. Branch

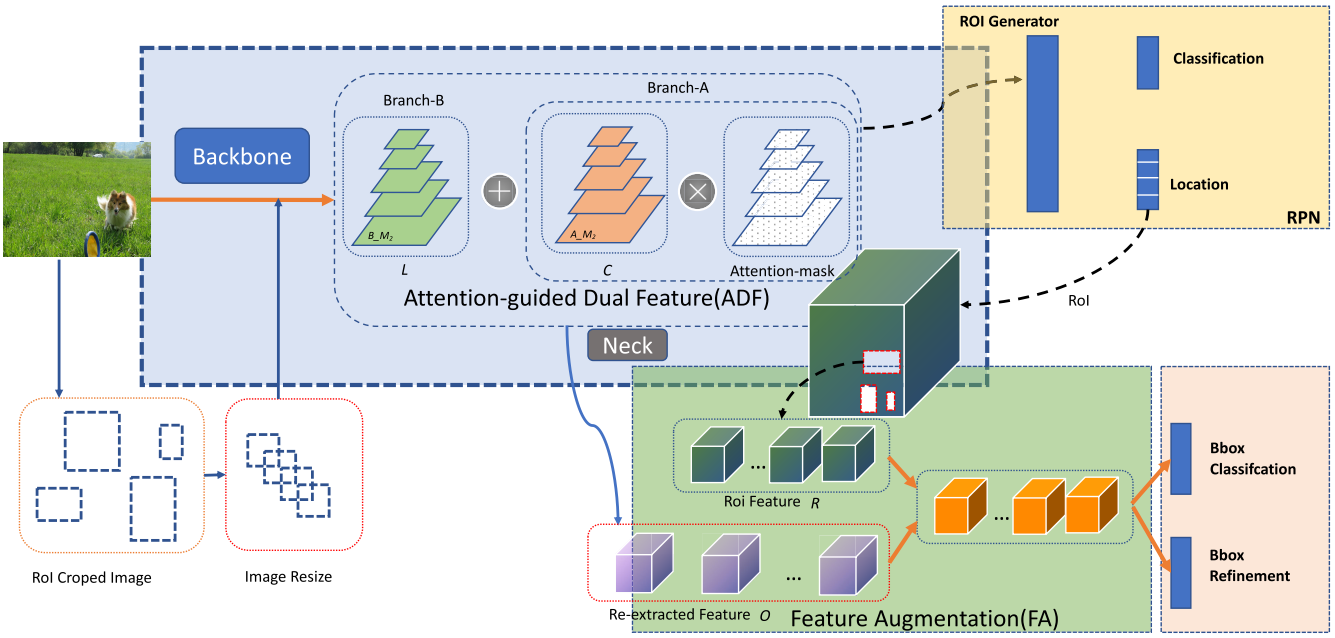


FIGURE 2. An overview of Att-FPA. The pipeline of our method is illustrated. The light blue rectangle shows the architecture of Attention-guided Dual Feature(ADF) model. The layers of the feature map L in branch A and the layers of feature map C in branch B are generated by the backbone (resnet [37], vgg [38] and so on). C is masked with the attention mask and jointed with L as the output of FPN. The red rectangles in dash line show the generate process of RoI image. The green rectangle shows the process of the Feature Augmentation. The features R are augmented by O and are used for serving the RCNN.

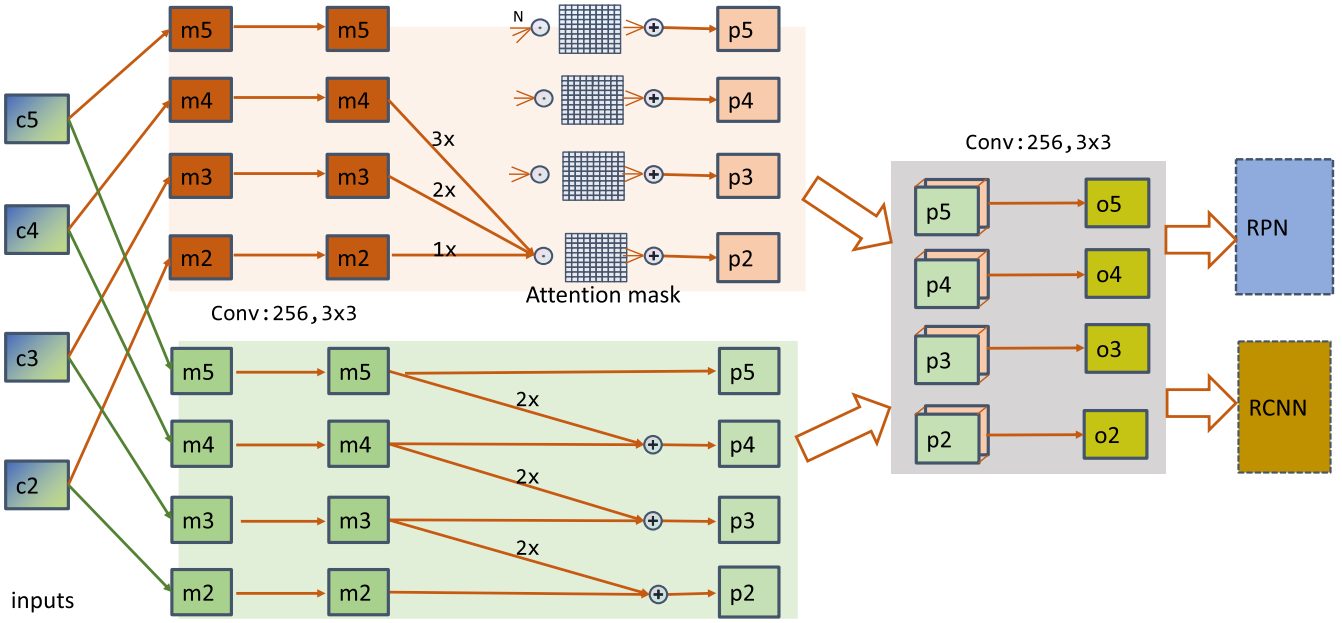


FIGURE 3. The Illustration of Dual Feature Pyramid Module. The light green block denotes the branch *B* carries more information on localization. The light beige block denotes the branch *A* is guided by the attention mask. The light grey block indicates the combined procession of the two branches. 1×, 2×, 3× indicate one, two, three times upsampling. The feature maps pyramid $\{O_2, O_3, O_4, O_5\}$ with the same channels are the output of the dual feature pyramid.

A is operated on 1×1 convolution filters and masked by the attention mask, which is perceived on localization. The 1×1 convolution filters are more sensitive for localization. It is mainly served for the localization sub-task. Branch B is operated with 3×3 convolution filters, which focuses on the semantic information. The 3×3 convolution filters are more suitable for classification. It is mainly served for the classification sub-task. In neck module, the procedure of output can be formulated as:

$$Output_i = Conv_i\big(Concat(\Gamma_i, B_M_i)\big). \tag{1}$$

where the Γ_i define as:

$$\Gamma_i = \sum_{j=i+1}^{N} \big(N_i(U_j(A_M_j)) \cdot (Att_i + \lambda)\big) + A_M_i \tag{2}$$

In (1), B_M, $\Gamma_i \in \mathbb{R}^{C,H,W}$, C, H, W denote channels, height and width of feature maps, respectively. In branch A, the Γ_i denotes the feature map which is guided by the attention mask map. The branches A and B are concatenated and operated on convolutional filters. In this paper, the bottom-up hierarchical feature fusion is used. In Eq.(2), N is the number of combined layers which are used for feature fusion. The current feature layers combine the features of the neighborhood to capture the semantic information. The high limit of N is the number of FPN layers. Upsampling (U_j) and normalizing (N_i) is performed on A_M_j and to fit in with the scale of A_M_i. λ is the hyper-parameter to keep the background feature information. A bigger *lambda* indicates a larger proportion of background information. Att_i is the attention mask map which is generated by the attention module. The detail of the module is shown in Fig. 3: the light green block indicates the branch B carries more information on localization. The light beige block denotes branch A is guided by the attention mask. The light grey block presents the combined procession of the two branches. In Fig. 3, 1×, 2×, 3× indicate the one, two, three times of up-sampling. The feature maps pyramid $\{O_2, O_3, O_4, O_5\}$ with same channels are the output of dual feature pyramid. The detailed steps of dual feature pyramid are described in Algorithm 1.

C. ATTENTION-GUIDED

In the architecture of object detection, the feature map of an image is generated from backbone. However, the pseudo objects and complex background have a high response in the feature map and which are easily to be detected. In this paper, we import an attention mask, the high response of pseudo objects is restrained and the response of foreground is enhanced. The attention mask is introduced to strength the ability of discrimination between foreground objects and background. The attention mask is learned by CNNs which is constrained by the ground truth mask. The ground truth boxes of manual annotation are used for generating the ground truth mask. The attention mask captures the geometrical information and provides meaningful differential features between the foreground objects and background.

As illustrated in Fig. 4, the left chart shows the details of our attention modules which are simple and effective. Att_i is generated by the attention modules due to the feature map $\{C_2, C_3, C_4, C_5\}$. The right image shows the process of the ground truth mask Ψ_i. It is a much easier format to generate due to the ground truth boxes. The ground-truth boxes are

Algorithm 1 Dual Feature Pyramid Algorithm

Input:
$\{C_2, C_3, C_4, C_5\}$ is input of the feature maps pyramid layers
Output:
$\{Output_i\}$ is output of the feature maps pyramid layers
Step 1: **for** each i ∈ [2, s] **do**
$L_i \leftarrow Conv_cls(C_i)$: operated on 3 × 3 convolution filters
$A_i \leftarrow Conv_reg(C_i)$: operated on 1 × 1 convolution filters
$Att_i \leftarrow Conv_att(C_i)$: operated on 1 × 1 convolution filters
end for
Step 2: **for** each i ∈ [2, s-1] **do**
$L_i \leftarrow L_i + \text{Upsample}(L_{i+1})$
$A_i \leftarrow A_i + \text{Upsample}(A_{i+1})$
end for
Step 3: **for** each i ∈ [2, s] **do**
$Output_i \leftarrow L_i + A_i \cdot (\lambda + Att_i)$
end for
Return: Att, Output

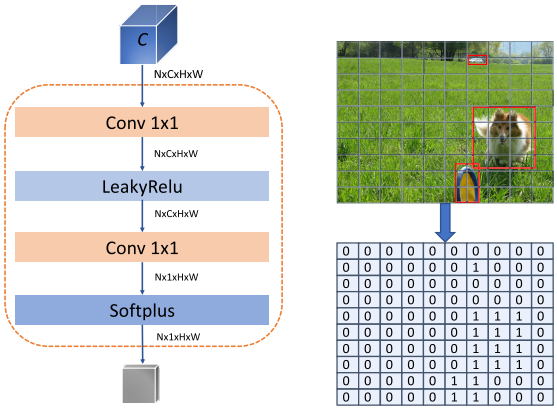


FIGURE 4. The Illustration of Attention Module. The left chart is attention modules which are simple and effective. It is mainly combined with convolutions operation. The amount of the feature map layers is decreased from C to 1. The right image shows the process of the ground truth mask Ψ_i. Once the coordinate of the feature map located in the ground truth, the value of mask is set to 1.

manually labeled. It is provided in Ms COCO label file [32]. The ground-truth mask is automatically generated due to these ground-truth boxes during the training process. The definition of Ψ_i is shown as follows:

$$\Psi_i = \begin{cases} 1 & (x_i, y_i) \in \mathbb{B}_i \\ 0 & (x_i, y_i) \notin \mathbb{B}_i \end{cases} \tag{3}$$

In (3), (x_i, y_i) denotes the coordinate of the feature maps at level i, $\mathbb{B}_i$ describes the corresponding ground truth bounding box sets. In this paper, ∈ means the coordinate located in the box. Att_i is constrained by Ψ_i, the optimization function is as follows:

$$\mathcal{L}_{att} = -\omega[\Psi_i \cdot \log\sigma(Att_i) + (1 - \Psi_i) \cdot \log(1 - \sigma(Att_i))] \tag{4}$$

$\mathcal{L}_{att}$ is the binary cross entropy with logits. The loss of RPN is defined as follows:

$$\mathcal{L}_{rpn} = \lambda_1 \cdot \mathcal{L}_{cls} + \lambda_2 \cdot \mathcal{L}_{loc} + \lambda_3 \cdot \mathcal{L}_{att} \tag{5}$$

In (5), the $\lambda_1, \lambda_2, \lambda_3$ are used for keeping trade-off between the loss of RPN.

D. FEATURE AUGMENTATION

In the recent work, feature augmentation focuses on the front parts, mainly on backbone and neck, such as adaptive fusion, residual feature augmentation, 3D convolution filters, context feature extraction and others. However, in two-stage framework, the bounding boxes (predicted by RPN) are used as the input of RCNN. The object features are extracted by RoI pooling layer. In this paper, it is regarded as $\mathcal{F}_{\mathcal{ROI}}$, one of the drawback is that the feature lack of perceiving: low-resolution inputs lack of information and high-resolution inputs fall into loss of information. In the framework of Faster-Rcnn, the feature is extracted from one single layer due to its scale. In this paper, we propose feature re-extraction and feature fusion in RoI feature extracting stage for feature augmentation. First of all, the predicted bounding boxes (predicted by RPN) with high score are selected. The input image is cropped based on the selected bounding boxes. The total amount of selected bounding boxes occupy a small proportion of samples in RPN, which is denoted as S. The cropped images are resized into the same tiny size which is counted as the normalization. As the inputs, the resized images are fed into the shared backbone and necks again for feature re-extraction. The re-extracted features are defined as $\mathcal{F}_{\mathcal{R}}$:

$$\mathcal{F}_{\mathcal{R}} = f(I_{crop}; \theta) \tag{6}$$

In (6), I_{crop} denotes the resized image, θ denotes the parameter of the model. Due to the reason that the proposal can be seen as one object, the ground truth bounding boxes are the size of the resized image.

$\mathcal{F}_{\mathcal{R}}$ is fed into RoI pooling layer to re-extract the feature. The features are denoted as $\mathcal{F}_{\nabla}$. The process of feature fusion and augmentation is calculated by:

$$\mathcal{F}_{\mathcal{AUG}_i} = Conv\Big(Concat\big(\sum_{j=1}^{M}(\mu_i^j \cdot \mathcal{F}_{\mathcal{ROI}}{}^j), \sum_{j=1}^{M}(\varrho_i^j \cdot \mathcal{F}_{\nabla}{}^j)\big)\Big) \tag{7}$$

In (7), M denotes the number of RoI layers. $\mathcal{F}_{\nabla}$ and $\mathcal{F}_{\mathcal{ROI}}$ are fused with the learning parameters by rescaling. After that, the features are augmented by a concatenated operation. The augmented features are used for multi-classification and localization improvements.

TABLE 1. The result of Att-FPA. The precision of Att-FPA on MS COCO *val*2017 dataset and ResNet and ResNext network are used and Faster-Rcnn architectures as a baseline. All the experiments using single model without any bells and whistles. The symbol ↑ denotes the relative improvement. The symbol * means the results come from the paper. The symbol † means the results come from MMdetection [15].

Method	Backbone	schedule	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLOv3_320†	DarkNet-53	23x	27.9	49.2	28.3	10.5	30.1	43.8
YOLOv3_416†	DarkNet-53	23x	30.9	52.9	32.1	14.4	33.5	44.9
YOLOv3_608†	DarkNet-53	23x	33.4	56.3	35.2	19.5	36.4	43.6
SSD_300†	VGG16	10x	25.6	43.8	26.3	6.8	27.8	42.2
SSD_512†	VGG16	10x	25.6	43.8	26.3	6.8	27.8	42.2
FCOS†	Resnet-50	1x	36.6	56.0	38.8	21.0	40.6	47.0
Faster-Rcnn†	Resnet-50	1x	37.4	58.1	40.4	21.2	41	48.1
Faster-Rcnn†	Resnet-101	1x	39.4	60.1	43.1	22.4	43.7	51.1
Faster-Rcnn†	X101-32x4d	1x	41.2	62.1	45.1	24.0	45.5	53.5
Faster-Rcnn†	X101-64x4d	1x	42.1	63.0	46.3	24.8	46.2	55.3
GA [32]†	X101-64x4d	1x	43.9	63.7	48.3	25.3	48.2	57.7
PA [22]*	Resnet-50	1x	37.5	58.6	40.8	21.5	41.0	48.6
PConv [21]*	Resnet-50	1x	38.5	59.9	41.4	-	-	-
AugFPN [16]*	Resnet-50	1x	38.7	61.2	41.9	24.1	42.5	49.5
AugFPN*	ResNet-101	1x	41.5	63.9	45.1	23.8	44.7	52.8
AugFPN*	ResNext-101-32x4d	1x	41.9	64.4	45.6	25.2	45.4	52.6
Att-FPA	Resnet-50	1x	40.5(↑3.1)	61.7	44.2	25.4(↑4.2)	44.0(↑3.0)	53.0(↑4.9)
Att-FPA	Resnet-101	1x	41.8(↑2.4)	62.6	45.6	25.8(↑3.4)	45.9(↑2.2)	54.9(↑3.7)
Att-FPA	ResNext-101-32x4d	1x	43.8(↑3.1)	64.7	48.2	27.6(↑4.2)	47.7(↑3.0)	57.4(↑4.2)
ADF+FCOS	Resnet-50	1x	38.5(↑1.9)	58.0	40.3	23.8(↑2.8)	42.5(↑1.9)	50.1(↑3.1)
Att-FPA+GA	ResNext-101-64x4d	1x	46.4(↑2.5)	66.1	50.9	29.2(↑2.9)	50.1(↑2.1)	60.0(↑2.7)

TABLE 2. Ablation study on Ms COCO val datasets. ADF represents attention-guided dual fpn which contains attention modules and dual feature pyramid. FA denotes the feature augmentation component.

Method	BaseNet	ADF	FA	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster-Rcnn	Resnet-50			37.4	58.1	40.4	21.2	41	48.1
Att-FPA	Resnet-50	✓		38.7(↑1.3)	59.7	41.7	22.9	42.3	49.7
Att-FPA	Resnet-50		✓	39.2(↑1.8)	60.1	42.9	23.7	42.8	51.5
Faster-Rcnn	Resnet-101			39.4	60.1	43.1	22.4	43.7	51.1
Att-FPA	Resnet-101	✓		40.2(↑0.8)	60.9	44.0	23.1	44.3	52.7
Att-FPA	Resnet-101		✓	40.7(↑1.3)	62.1	44.1	24.2	44.2	54.8

IV. EXPERIMENTS

A. DATASETS AND EVALUATION

In the paper, we do experiments on Ms COCO datasets with 80 categories which have been widely acknowledged. We trained Att-FPA on *train*2017 datasets with 118K images. The evaluation is performed on *val*2017 dataset with 5k images. The results of our paper are reported by standard COCO-style metrics. Average Precision (AP) are averaged over multiple Intersection over Union (IoU) values and all categories.

B. THE TRAINING DETAILS OF ATT-FPA

The same as other methods, our experiment images are resized to have a minimum dimension of 800 and maximum dimension of 1333. We train Att-FPA on 8 GPUs (Tesla V100) for 12 epochs and one image per GPU. The images in feature augmentation are resized to 128. It is small but enough for feature augmentation. In this paper, the features of resized image are built on C_1, C_2 [16] and the amount of calculation is reduced greatly. The data augmentation only contains a random flip operation. The weights of all models are initialized from pre-trained models on ImageNet [33]. The learning rate of training is 0.01, which is warmed up in 1000 iterations with the ratio of 0.3. It is decreased to 0.001 at epoch 8 and 0.0001 at epoch 11. In Eq.(2), λ is set to be 1.0. It means that the foreground objects have twice response than the background. The λ_1, λ_2, λ_3 are set to be 1.0, 1.0, 0.05 respectively. We evaluate our method on Ms COCO object detection datasets. we use {0.5, 1.0, 2.0} aspect ratios one scales {8} for the anchors. The output size of RoI layer is adjusted from 7 to 8, which is matched with the feature augmentation stage.

C. EXPERIMENTAL RESULTS

In this section, we explore the result of Att-FPA on Faster-Rcnn. In Table. 1, we summarize the mAP for each module of Att-FPA. As the table shows, Att-FPA achieves a good performance on Ms COCO *val*2017 datasets.

It is obvious that the Att-FPA has 3.1% improvement with the combination of the two components without any ablation. Moreover, Att-FPA has achieved a significant improvement on small and large objects (4.2% and 4.9% respectively) than medium size objects (3.0%). Att-FPA supplies and enhances the feature expression for both small and large object. Fig. 5 shows the training process between Faster-Rcnn (red) and Att-FPA (blue). During the 12 epochs, the proposed method improves the performance significantly after epoch 8. We further train Att-FPA with different backbone networks on the Ms COCO datasets. As Table. 1 shows, Att-FPA still has an

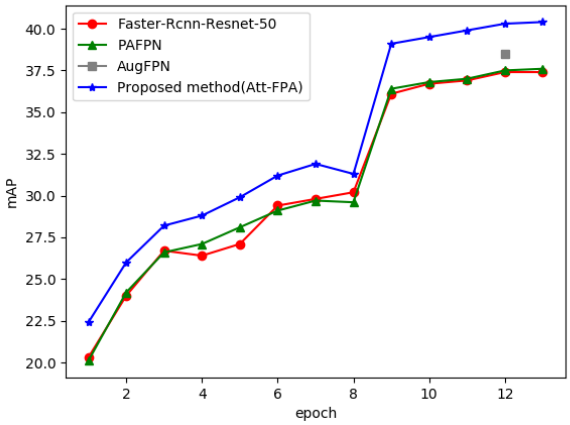


FIGURE 5. The Illustration of mAP at different training epochs. The figure describes the mAP during the training process between Faster-Rcnn (red) and the proposed method (blue). The proposed method improves the performance significantly after epoch 8.

TABLE 3. Ablation study on FA with multi-test experiment.

Model name	Scale mode	mAP
Faster-Rcnn	single-scale	37.4
Faster-Rcnn	multi-scale	38.7 (↑1.4)
FA	single-scale	39.4
FA	multi-scale	40.3 (↑0.9)

TABLE 4. The result of Att-FPA on VOC2007 + 2012. In Pascal VOC2007 + 2012 dataset. The Att-FPA achieves 9+% mAP promotion.

Model name	BaseNet	mAP
Faster-Rcnn	Resnet-50	79.5
Att-FPA	Resnet-50	88.6

outstanding performance on large models such as ResNext-101-32x4d, ResNext-101-64x4d, which have a promotion about 2%∼3%, respectively. Take ResNext-101-32x4d as an example, the mAP has a promotion by 3.1%, which is the same as Resnet-50. This demonstrates that the proposed Att-FPA has a great presentation of features and is robust to a large model. It also works on small and large objects and obtains a promotion experimentally. The proposed methods also performed evaluation on Pascal VOC datasets [35], which is a common object detection datasets. As Table. 4 shows, the Att-FPA achieves 9+% AP promotion on the common object detection datasets Pascal VOC. GA, PA and AugFPN do not provided the results of VOC. The comparison is focus on Ms COCO in the following work.

Although the proposed Att-FPA achieves a better result on Ms COCO. Table. 5 showing our model complexity for Att-FPA, we use the same backbone, neck and prediction

TABLE 5. Comparation between Faster-Rcnn and Att-FPA on computational complexity.

Module	BaseNet	ADF	FA	AP	Params ratio	FLOPs ratio
Faster-Rcnn	Resnet-50			37.4	1x	1x
Att-FPA	Resnet-50	✓		38.7	1.1x	1.08x
Att-FPA	Resnet-50		✓	39.2	1.08x	1.76x

TABLE 6. Comparation between Faster-Rcnn and Att-FPA on processing time.

Module	BaseNet	ADF	FA	AP	Mem(GB)	Inf time(fps)
Faster-Rcnn	Resnet-50			37.4	4.0	21.4
Att-FPA	Resnet-50	✓		38.7	4.5	19.67
Att-FPA	Resnet-50		✓	39.2	15.6	12.06

network. The proposed ADF module brings about 10% increment of the base but has 1.3% mAP improvement. The proposed FA module bring about 70% complexity increments of the base and has 1.8% mAP improvements. We measure the inference time of Att-FPA in Table. 6. The memory access cost has 12.5% increasing and the inference time cost is about 50.8ms. The ADF has a frame rate of about 20 frames per second(fps). The proposed FA module has a large memory access of about 16 GB, and the inference time cost is as high as 80.3ms. The ADF and FA modules are conveniently to be implemented in other models.

D. COMPARISON WITH STATE OF THE ART

Att-FPA is compared with several recent state-of-the-art (SOTA) feature augmentation methods. The last five rows in Table. 1 show the comparison. The results of state-of-the-art come from the corresponding paper. In a similar way, our results are carried out without bells and whistle (such as multi-scale training and testing, GN, and DCN). Att-FPA outperforms the existing approaches on feature fusion and augmentation. The results demonstrate that Att-FPA has a better feature to perceive on small objects and large objects. In this paper, the proposed method of boosting feature perceiving for object detection has a great and steady performance.

E. VISUALIZATION OF MODULE

In Fig. 6, the first row images are the heat map of the Faster-Rcnn-Resnet101 and Att-FPA, the second row images are the corresponding result of detection. More results are shown in Fig. 8, It is observed that the feature map of Att-FPA clearly distinguishes the objects from the background. The most highlighted areas are located in foreground objects. Besides, the information of background is also captured. In Fig. 7, the details of the comparison are shown. The red arrow shows some of the pseudo objects are filtered with Att-FPA.

F. ABLATION STUDIES

Att-FPA is employed in the model to capture more information for boosting feature perceiving. In this section, the performance of the components are verified. As shown in Table. 2, the components ADF and FA have a great performance. In the proposed framework, the dual feature pyramid and attention-guided would be trained jointly in back-propagation. Attention-guided dual FPN (ADF) modules improve the performance by 1.3%. Meanwhile, employing feature augmentation (FA) module individually outperforms the baseline by 1.8%. For the feature augmentation, there is an assumption that the performance would be ablated by

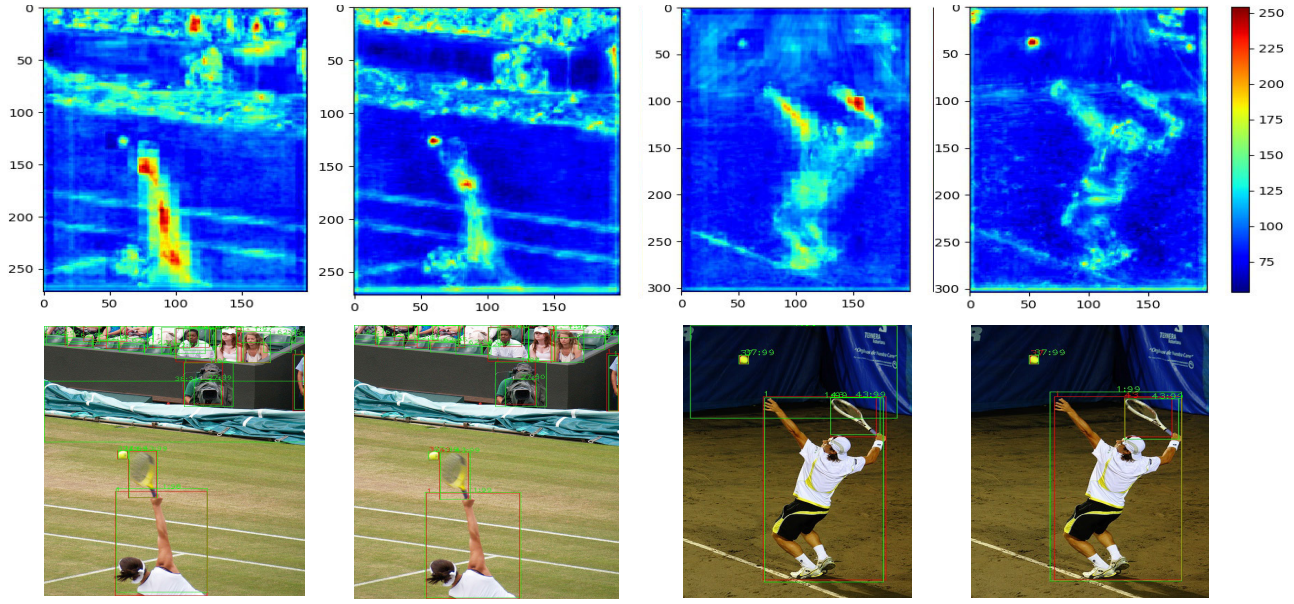


FIGURE 6. The Output of detector.The heat map of the Faster-Rcnn-Resnet101 and Att-FPA are shown at the first row, the second row are the corresponding result of detection. The first and third column images are the result of Faster-Rcnn-Resnet101. The second and fourth column images are the result of Att-FPA. It is observed that the foreground (tennis ball) has a high response and less FP.



FIGURE 7. The details of comparison on detection bounding boxes. The first and third column images show the result of Faster-Rcnn-Resnet101. The second and fourth column images show the result of Att-FPA. The red arrow shows the FP area.



FIGURE 8. More results of Att-FPA compared with Faster-Rcnn.

multi-scale test. In this paper, we do some experiments on the assumption. The size of test images are transformed over the scales as {(1333, 800), (1000, 600), (1600, 1000), (900, 600), (1200, 800)}. And the test images are randomly flipped. As Table. 3 shows, the multi-scale test on Faster-Rcnn improve the performance by 1.4%. Although the promotion is lower than Faser-Rcnn, FA with multi-scale still improves the performance by 0.9%, which is an improvement.

TABLE 7. Ablation study on FA with multi-test experiment.

Module	BaseNet	λ	map	AP_S	AP_M	AP_L
ADF	Resnet-50	0	38.4	22.5	41.9	49.3
ADF	Resnet-50	1	38.7	22.9	42.4	50.0
ADF	Resnet-50	2	38.1	21.9	42.1	48.8
ADF	Resnet-101	0	40.0	23.6	44.0	52.2
ADF	Resnet-101	1	40.2	23.1	44.3	52.7
ADF	Resnet-101	2	40.1	22.7	44.2	52.2

In Eq. (2), λ is defined for keeping the background with a sensitive reflect. λ in Eq.(2) is set as zero, which means the background feature information will be covered up in extremes. λ equals to one indicates the foreground object features have twice more effects than the background area. The bigger λ indicates a lower ability of discrimination between foreground object and background. Table. 7. shows that λ equals to one achieve better result than others, which is comprehensible intuitively.

V. CONCLUSION

The performance of object detection has been significantly improved due to the various design of deep convolutional neural networks. However, the backbone network is designed for the image classification tasks, which have conflicts with the tasks on object detection [36]. A detection-aware architecture is an important research direction [39]. Differently from [36]. In this paper, our research focus on the detection-aware neck by enhancing the feature perceive for object detection. We present Att-FPA, a new feature perceive and augmentation architecture, which is designed specifically for object detection. Att-FPA take the feature of subtasks of classification and localization into account at the same time. An attention mask with new constraint information is imported. Based on this, a novel dual feature pyramid network is designed to serve for classification and localization tasks. Besides, information is re-extracted from backbone and neck for RoI feature augmentation. Because of the new design, the indistinguishability between features at foreground and background is eliminated. The conflicts between classification and object detection tasks are addressed. And the ability of feature representation in RoIs is enhanced. The ablation experiments show the Att-FPA improve the performance effectively. The proposed components improve the performance without ablated each other. Att-FPA achieves excellent performance on feature presentation and real-world applications experimentally demonstrate the effectiveness and feasibility of the proposed approach.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.* vol. 25, 2012, pp. 1097–1105.

[2] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Finding tiny faces in the wild with generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 21–30.

[3] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 28, 2019, doi: 10.1109/TPAMI.2019.2956516.

[4] Y. Li, J. Li, W. Lin, and J. Li, "Tiny-DSOD: Lightweight object detection for resource-restricted usages," 2018, *arXiv:1807.11013*. [Online]. Available: http://arxiv.org/abs/1807.11013

[5] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Light-head R-CNN: In defense of two-stage object detector," 2017, *arXiv:1711.07264*. [Online]. Available: http://arxiv.org/abs/1711.07264

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[7] X. Yu, Y. Gong, N. Jiang, Q. Ye, and Z. Han, "Scale match for tiny person detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1257–1265.

[8] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[9] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 821–830.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*. [Online]. Available: http://arxiv.org/abs/1506.01497

[11] W. Liu, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[13] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[14] A. Womg, "Tiny SSD: A tiny single-shot detection deep convolutional neural network for real-time embedded object detection," in *Proc. 15th Conf. Comput. Robot Vis. (CRV)*, 2018, pp. 95–101.

[15] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, and Z. Zhang, "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*. [Online]. Available: http://arxiv.org/abs/1906.07155

[16] C. Guo, B. Fan, Q. Zhang, S. Xiang, and C. Pan, "AugFPN: Improving multi-scale feature learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12595–12604.

[17] A. Jamaludin, T. Kadir, and A. Zisserman, "SpineNet: Automatically pinpointing classification evidence in spinal MRIs," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 166–175.

[18] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3, 2006, pp. 850–855.

[19] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.

[20] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9759–9768.

[21] X. Wang, S. Zhang, Z. Yu, L. Feng, and W. Zhang, "Scale-equalizing pyramid convolution for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13359–13368.

[22] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[23] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7036–7045.

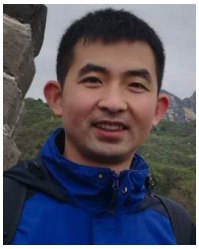
[24] Y. Liu, "CBNet: A novel composite backbone network architecture for object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11653–11660.

[25] G. Song, Y. Liu, and X. Wang, "Revisiting the sibling head in object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11563–11572.

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473

[27] M. A. Islam, S. Jia, and N. D. B. Bruce, "How much position information do convolutional neural networks encode?" 2020, *arXiv:2001.08248*. [Online]. Available: http://arxiv.org/abs/2001.08248

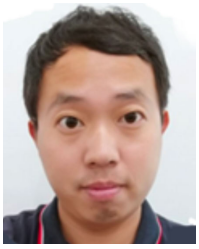
[28] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.
[29] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," 2019, *arXiv:1906.05909*. [Online]. Available: http://arxiv.org/abs/1906.05909
[30] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3286–3295.
[31] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[32] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2965–2974.
[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[34] T.-Y. Lin, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.
[35] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[36] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "DetNet: A backbone network for object detection," 2018, *arXiv:1804.06215*. [Online]. Available: http://arxiv.org/abs/1804.06215
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[39] X. Wu, D. Sahoo, and S. C. H. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, vol. 396, pp. 39–64, Jul. 2020.



JINGWEI LIU received the M.S. degree in computer science from Peking University, in 2012. He is currently pursuing the Ph.D. degree with the Institute of Computing Technology, Chinese Academy of Sciences (ICT-CAS). His research interest includes computer vision: image classification, object detection, object segmentation algorithms and applications, and self-supervised visual feature learning.



YI GU received the B.E. degree in mechanical engineering and automation from the Beijing University of Posts and Telecommunications, Beijing, China, in 2011, and the M.S. and Ph.D. degrees in electrical engineering from the Ritchie School of Engineering and Computer Science, University of Denver, Denver, CO, USA, in 2014 and 2020, respectively. Her research interests include computer vision, power system optimization, machine learning, deep learning, and data analysis.



SHUMIN HAN received the M.S. degree in computer science from Peking University, in 2012. He is currently a Senior Engineer with Baidu Online Network Technology (Beijing). His research interest includes computer vision: image classification, object detection, object segmentation algorithms and applications, and self-supervised visual feature learning.



ZHIBIN ZHANG (Member, IEEE) is currently an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He also leads a research group working on system research. Since 2014, he has been focusing on big data systems and leads a high-performance graph computing system SQLGraph and a neural network compiler MLCC. He has published more than 20 papers on top conferences, including INFOCOM, ICNP, DEX, and JSAC. His research interests include graph systems, streaming processing systems, and machine learning systems. He has received the Best Paper Award of ICNP 2012 and DEX 2019.



JIAFENG GUO (Member, IEEE) received the B.E. degree in computer science and technology from the University of Science and Technology of China, Hefei, China, in 2004, and the Ph.D. degree in computer software and theory from the University of Chinese Academy of Sciences, Beijing, China, in 2009. He is currently a Researcher with the Institute of Computing Technology, Chinese Academy of Sciences, a Professor with the University of Chinese Academy of Sciences, and a PI with the Beijing Academy of Artificial Intelligence. He is also the Chief-Vice Director of the CAS Key Laboratory of Network Data Science and Technology. He has worked on a number of topics related to web search and data mining, including query representation and understanding, learning to rank, and text modeling. His current research interests include neural models for information retrieval (Neural IR) and natural language understanding. He has won the Best Paper Award in ACM CIKM 2011, the Best Student Paper Award in ACM SIGIR 2012, and the Best Full Paper Runner-Up Award in ACM CIKM 2017.



XUEQI CHENG (Senior Member, IEEE) is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (ICT-CAS) and the Director of the Research Center of Web Data Science & Engineering (WDSE), ICT-CAS. He has published more than 100 publications in prestigious journals and conferences. His main research interests include network science, web search and data mining, big data processing, and distributed computing architecture.