Feature Learning Improved by Location Guidance and Supervision for Object Detection

BINGYING LI\textsuperscript{1,2}, JIALE XIONG\textsuperscript{1,3}, XIANG FU\textsuperscript{1,3}, JIEXIAN ZENG\textsuperscript{1,2,3}, AND LU LENG\textsuperscript{1,3,4}, (Member, IEEE)

\textsuperscript{1}Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition, Nanchang Hangkong University, Nanchang 330063, China
\textsuperscript{2}Science and Technology College, Nanchang Hangkong University, Nanchang, Gongqingcheng 332020, China
\textsuperscript{3}School of Software, Nanchang Hangkong University, Nanchang 330063, China
\textsuperscript{4}School of Electrical and Electronic Engineering, College of Engineering, Yonsei University, Seoul 120749, Republic of Korea

Corresponding author: Lu Leng (leng@nchu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61763033, Grant 61662049, Grant 61866028, Grant 61663031, and Grant 61866025.

\section*{ABSTRACT} In recent years, the single-stage detectors have been developed rapidly; however, compared with the multi-stage detectors, their detection precision is still relatively low. Single-stage detectors and multi-stage detectors are analyzes and compared in detail in this paper, which reveals that single-stage detectors suffer from some problems, including feature loss and inaccurate feature extraction. Therefore, this paper proposes a novel detection model, dubbed Optimized Network (OptNet), to alleviate these deficiencies. OptNet consists of three modules: pyramid of attention features, feature alignment and consistency supervision (CS). The pyramid of attention features, based on feature pyramid networks (FPNs), introduces a novel branch named attention FPN (AtFPN), which aggregates the multi-layer features of the backbone network and optimizes the object features by using lightweight attention modules. AtFPN alleviates the loss of the feature pyramid information and the blocking of feature transmission between adjacent layers. Meanwhile, it provides global information for the model. The feature alignment module aligns the anchor box to the feature by using the object location information to guide the network to extract precise object features. Finally, CS accelerates network optimization and reduces semantic differences between the features on different layers. In the detection stage, OptNet optimizes the prediction of the model with the first detection result to improve the accuracy. Experiments on the MS COCO 2017 dataset demonstrate that OptNet yields significant improvement in the detection precision.

\section*{INDEX TERMS} Object detection, feature alignment, multiple detection, consistency supervision.

\section*{I. INTRODUCTION} Object detection is one of the basic fields of computer vision. Its core task is to identify and localize the objects of interest in the images. With the rapid development of deep learning in recent years, a number of state-of-the-art detectors have been proposed based on deep learning. These object detection algorithms can be briefly divided into two categories: multi-stage detection \cite{1}–\cite{5} and single-stage detection \cite{6}–\cite{9}. The core idea of multi-stage object detection is to accurately determine the object position and category by multiple detection; while single-stage object detection relies on a single fully convolutional networks to predict the localization and classification of the object. Therefore, in terms of detection accuracy, multi-stage detectors can obtain high accuracy through multiple refined optimization of object localization and features. However, multi-stage detectors typically require high computational complexity, which degrades the detection efficiency. Single-stage detection models are commonly light-weight and simple, and they can detect object quickly. Therefore, single-stage detection models have better prospect in real-time applications.

In order to improve the accuracies of single-stage detectors, many researchers have analyzed the challenges degrading the single-stage detection models, and proposed the corresponding solutions. Nas-FPN \cite{10} used a neural architecture search (NAS) to find the best feature extraction network in a specific search space. Shrivastava \textit{et al}. \cite{11}...
FIGURE 1. Overall pipeline of feature pyramid network and its deficiencies: 1) object feature loss at deference pyramid level; 2) features only focus on adjacent features, leading to blocked information transmission; 3) feature pyramid features have semantic differences. Number in each parallelogram box represents feature channel number.

FIGURE 2. Anchor box does not match object features. Gray represents convolution kernel and black represents features obtained by convolution.

paid more attention to hard samples to enhance the performance of the detector. Zhu et al. [12] automatically assigned the objects to specific layers for detection. These methods improved the precision of the single-stage detector, but compared with the multi-stage detection, the single-stage detection algorithm still has the following disadvantages.

A. LIMITATIONS OF LOCAL FEATURES

A convolutional neural network (CNN) only considers the local information of the object, which is formed by a stack of multiple convolutional layers, while neglects global information. Local information refers to the feature extracted from the local area in the image and can accurately describe the basic object characteristics; while global information refers to the overall attribute of the image. For some similar objects, global information can better describe the objects, so the reasonable usage of global information and local information is conducive to object detection.

B. SEMANTIC GAPS BETWEEN FEATURES

Each layer of the feature pyramid has its specific semantics. As shown in Figure 1, most object detection algorithms distribute the objects to different feature layers for detection, and each layer only detects the objects within its specific range. Therefore, different feature layers pay attention to different object information. However, to obtain better object features, feature pyramid networks (FPN) [13] uses high-level features to optimize low-level features. These operations do not consider the semantic differences between the different features of the feature pyramid. Although FPN is simple and effective, it is suboptimal. In addition, FPN does not make sure that each layer of feature maps only detects the objects within a specific range.

C. MISALIGNMENT BETWEEN ANCHOR BOX AND FEATURE

As shown in Figure 2, the black feature extracted by convolution covers only a part of the object features, so the anchor generated by this feature is not aligned with the object features. Multi-stage detection typically filters and localizes the object box several times, and then extracts the features of the region where the object exists. Therefore, multi-stage detection alleviates the misalignment between the anchor box and feature to a certain extent. However, single-stage object detectors typically directly use the object features extracted by CNN to predict the category and location of the anchor box, resulting in a relatively lower detection precision.

Based on the above discussions, the optimized network (OptNet) is proposed in this paper, which designs three different modules to address the aforementioned problems. Our main contributions can be summarized as follows.

- A novel feature pyramid structure, dubbed attention FPN (AtFPN) is proposed to obtain representative object features through multiple sampling of multi-scale features. It compensates for the deficiency of the traditional feature pyramid.
- Furthermore, consistency supervision (CS) module is used for the supervising learning of multi-scale features of objects to eliminate the semantic differences between features. This module uses the real information of the object to supervise the learning of the network, so each layer of feature maps only detects the object in a specific range.
- To achieve alignment between anchor box and feature, the location offset obtained by CS is used in the feature alignment module. The position information of anchor is employed to guide the network feature extraction.

II. RELATED WORK

In the field of computer vision, the main task of neural networks is to obtain the optimal features of the detected objects. For this reason, a large number of neural network structures have been designed. Currently, there are three ways to obtain the optimal features of the object: (1) improve the network such as the backbone and FPN; (2) use the attention mechanism; (3) use the location or shape information of the object to optimize the feature map. Among these approaches, building a high-quality feature extraction network is the main
way to obtain the unique information of the object. To explore a better network structure, VGG [17] used multiple $3 \times 3$ convolutions instead of a large convolutional kernel. This operation increased the receptive field of the network while reducing the network parameters. GoogLeNet [18]–[21] proposed the inception network structure, which increased the width of the network and used small convolution kernels instead of large convolution kernels. GoogLeNet achieves better precision with a lower number of parameters. ResNet designed a residual module, which effectively expanded the depth of the neural network. The residual module added the input of the upper layer to the output of this layer. This structure accelerated the training of the network. DenseNet [22] created short paths from early layers to later layers to ensure maximum information transfer between different layers of the network. This network achieves better performance than ResNet with fewer parameters. GcNet [23] combined the idea of non-local [24] and SeNet [25] to obtain better global context information. SpineNet [26] used NAS strategies to explore a truly suitable backbone network for object detection that allowed for cross-scale connectivity between features, and the feature scales could be scaled up or down as needed.

Although an excellent backbone network can extract better object features, how to construct an optimal feature pyramid structure is also very important in the field of object detection. FPN is a highly important network structure that transfers the features from the high level to the low level through a top-down structure. This structure alleviates the problem of multi-scale object detection. PANet [27] designed a bottom-up structure on the basis of FPN that transmitted low-level features to high-level features to reduce the transmission distance of information. Libra R-CNN [28] aggregated the features of all layers of the feature pyramid and then used the aggregated features to refine the features of FPN. AugFPN [29] designed residual feature augmentation to enhance the high-level character.

Since fusion can effectively improve the complementary advantages of fused sources [30], [31], Chu et al. [32] proposed multi-layer feature fusion to obtain high-quality object features. Mask refined R-CNN [33] utilized the refined information extracted from the network to optimize features. ASFF [34] proposed an adaptive feature fusion method to eliminate the inconsistency of the feature pyramid. NAS-FCOS [35] used NAS technology to construct a feature pyramid that improved the performance of the detector. There are many other ways to use NAS technology to build feature pyramids, including MnasFPN [36] and NAS-FPNLite [36]. Different from the above methods, this paper redesigns the feature pyramid to make full use of shallow details and strengthens the information interaction among the features of the feature pyramid. This model alleviates the shortcomings of the original feature pyramid.

Attention is a network model constructed by imitating the mechanism of human visual attention. The core goal is to make the network pay attention to the information through the corresponding operation while suppressing or discarding the useless information. SKNet [37], as a dynamic selection network, allows each neuron to adaptively adjust its receptive field size according to the input multi-scale information. CBMA [38] uses a channel attention mechanism and spatial attention mechanism to make the network focus on important features and suppress interference features. Different from them, we design an attention module using global average pooling to obtain the context information of the object.

The location of an object in an image can reflect the range of the object feature in the feature map. However, due to the interference of various factors, the features within the object area are easily disturbed by the surrounding features. Therefore, using the object location or shape information to guide the network can help the network extract the object features more accurately. Deformable ConvNet [39], [40] proposed deformable convolution, which employed the object location information predicted by the network to guide the network to better extract object features. ThunderNet [41] uses the location information of the candidate box to guide the regional feature extraction module to obtain the features within its range. AlignDet [42] designed the ROIConv module, which uses the calculated offset information to guide the network to extract more accurate object features. Similar to the above methods, this paper uses the object location information predicted by the network as the input of the deformable convolution to guide the network to obtain better object features.

III. METHODOLOGY

OptNet is an object detector based on the RetinaNet framework, so we first review RetinaNet for object detection in Section III. A. Figure 3 shows the overall network structure of OptNet. By using the attention FPN module, feature alignment module and CS, OptNet alleviates the insufficient feature extraction in single-stage detection, as well as low classification and localization precision. In Section III. B, we explore how to enable the network to extract accurate object features, and introduce the design of the AtFPN in detail. In Section III. C, we describe how the feature alignment module alleviates the misalignment between the anchor box and the feature in single-stage detection. Section III. D specifies how OptNet uses CS to eliminate the effects of semantic differences.

A. REVIEW OF RetinaNet FOR OBJECT DETECTION

RetinaNet is an anchor-based single-stage object detector. It uses ResNet as the backbone network and FPN to construct feature maps of different scales. In the training stage, RetinaNet generates a large number of anchor boxes on each layer of the feature pyramid, which are used to determine the object location in the image. After that, the detector assigns objects to different layers of the feature pyramid according to the intersection over union (IoU) between the anchor box and the ground true box. Each layer of the feature map is connected to a prediction network for object classification and regression prediction.
The parameters of the prediction network are shared, which accelerates the detection speed. However, due to the limited number of objects in the image, most anchor boxes are invalid, so the numbers of positive and negative samples are imbalanced. To alleviate the impact of negative samples on the detector, RetinaNet leverages a novel classification loss function called focal loss that is based on cross-entropy loss. The loss function controls the proportion of positive and negative samples, prevents imbalance by restraining the weights of easy samples, and focuses the detector’s attention on hard samples. The focal loss is defined as follows:

$$FL(P_t) = -\alpha_t (1 - P_t)^\gamma \log(P_t)$$  \hspace{1cm} (1)

$P_t$ is the estimated probability for the class. The adjustment factor $(1 - P_t)^\gamma$ adjusts the contributions of hard and easy samples to loss, and the balance factor $\alpha_t$ controls the balance of positive and negative samples.

B. AtFPN

FPN is a lightweight and effective feature extraction network that improves the detection performance of the detector to a certain extent. However, FPN uses convolution force to reduce the channel number of feature maps, which makes the network lose part of the effective information. At the same time, in the feature fusion stage, features are fused layer by layer downward, so there is information transmission blockage between the features of the feature pyramid.

To solve these problems, this paper designs a novel feature pyramid network, dubbed AtFPN, as shown in Figure 3. Since FPN is lightweight and effective, its original structure is retained in the design of the feature pyramid, and then a novel feature extraction module attention feature is constructed on FPN. The attention feature module (AFM) is parallel to the FPN and shares the same feature maps. This module is mainly used to obtain more abundant object features. AFM consists of five steps: feature extraction, scaling, fusion, optimization and enhancement.

Figure 4 shows the overall network structure of the AFM. Compared with the FPN network, AFM uses a $3 \times 3$ convolution kernel to extract the features of the backbone network. This operation increases the receptive field of the network and simultaneously obtains richer semantic information. Then, the obtained multiple features are fused. Since lower-level feature maps have rich object details, the AFM first uses linear interpolation to sample higher-level features, adds them to lower-level feature maps. This operation is used to extract local features. After that, global max-pooling is carried out on $P_1$ to extract global information. Then, $1 \times 1$ convolutions optimize the global information. Finally, the global information is fused with $P_1$ to obtain AtFeature. To enhance the information communication between high-level and low-level features, AFM uses adaptive max-pooling to scale the AtFeature to a specific size. Then, these feature maps are fused with FPN features to obtain features $\{R_3, R_4, R_5\}$, as shown in Figure 3. Then, a $3 \times 3$ convolution is used to convolve the high-level features to obtain the features $\{R_6, R_7\}$. The employed attention module is similar to that of GcNet [23], but only convolution and global max-pooling are used to reduce the complexity of the network.

C. FEATURE ALIGNMENT

Two-stage detection identifies and locates the object twice. The detector first recognizes the approximate location of the object in the image, and then the optimized anchor box guides ROI pooling to extract object features. After that, the extracted features are used to classify and locate the anchor box for
Based on this discovery, a feature alignment module is designed to alleviate the misalignment between the anchor box and features in single-stage detection. As shown in Figure 5, two inputs of the feature alignment module are the classification feature map and regression feature map. Regression_I is the anchor box location offset predicted by the regression feature map at the first time. To obtain the location offset of the convolution kernel, Regression_I is convolved with a $1 \times 1$ convolution. Then, the predicted offset is input to the feature alignment module, which uses the predicted offset to guide the deformable convolution to better optimize the classification and regression feature map. In this work, we use a $3 \times 3$ deformable convolution to realize the alignment between the anchor box and feature map.

**D. CONSISTENCY SUPERVISION**

FPN uses multi-scale features of the backbone network to build a feature pyramid. To ensure that low-level features have rich semantic information, FPN uses the top-down method to transfer high-level features to low-level features. In the detection stage, the detector disperses the object to different layers of the feature pyramid for detection; however, each layer of the feature pyramid is more inclined to acquire the desired object feature. Therefore, the features of different scales contain different information, and there are significant semantic differences between them. This leads to a suboptimal feature pyramid generated by simple aggregation of multi-scale features.

To eliminate the influence of multi-scale features on the detector, OptNet uses the real label and ground true box of the object in the image to supervise the learning of the feature pyramid. This ensures that each layer feature of the feature pyramid only detects the object within a specific range. Therefore, the detector obtains the first detection results of Classification_I and Regression_I. Regression_I is the location offset of the anchor box, which guides the prediction subnet to obtain finer features. After that, we use repeated $3 \times 3$ convolution to convolve the obtained features.
TABLE 1. Comparison with state-of-the-art methods on COCO test-dev. "∗" indicates reimplementation results in MMDetection.

| Method         | Backbone   | AP   | AP_{50} | AP_{75} | AP_{S} | AP_{M} | AP_{L} |
|----------------|------------|------|---------|---------|--------|--------|--------|
| YOLOv2         | DarkNet-19 | 21.6 | 44.0    | 19.2    | 5.0    | 22.4   | 35.5   |
| YOLOv3         | DarkNet-53 | 33.0 | 57.9    | 34.4    | 18.3   | 35.4   | 41.9   |
| SSD300         | VGG16      | 23.2 | 41.2    | 23.4    | 5.3    | 23.2   | 39.6   |
| SSD512         | VGG16      | 26.8 | 46.5    | 27.8    | 9.0    | 28.9   | 41.9   |
| Faster R-CNN   | ResNet-101-FPN | 36.2 | 59.1    | 39.0    | 18.2   | 39.0   | 48.2   |
| Deformable R-FCN | Inception-ResNet-v2 | 37.5 | 58.0    | 40.8    | 19.4   | 40.1   | 52.5   |
| Mask R-CNN     | ResNet-101-FPN | 38.2 | 60.3    | 41.7    | 20.1   | 41.1   | 50.2   |
| Faster R-CNN∗  | ResNet-50-FPN | 36.2 | 58.5    | 38.9    | 21.0   | 38.9   | 45.3   |
| Faster R-CNN∗  | ResNet-101-FPN | 38.8 | 60.9    | 42.1    | 22.6   | 42.4   | 48.5   |
| RetinaNet*     | ResNet-50-FPN | 36.4 | 56.5    | 38.9    | 20.7   | 39.5   | 44.7   |
| RetinaNet*     | ResNet-101-FPN | 38.5 | 58.8    | 41.4    | 21.3   | 42.0   | 48.7   |
| OptNet         | ResNet-50-FPN | 37.5 | 57.8    | 40.4    | 21.5   | 40.4   | 46.4   |
| OptNet         | ResNet-101-FPN | 39.4 | 59.7    | 42.4    | 22.5   | 42.4   | 50.0   |

TABLE 2. Comparisons between RetinaNet and OptNet.

| Method      | Backbone          | AR_{max-1} | AR_{max-10} | AR_{max-100} | AR_{S} | AR_{M} | AR_{L} |
|-------------|-------------------|-------------|--------------|---------------|--------|--------|--------|
| RetinaNet   | Resnet50-FPN      | 31.5        | 50.7         | 54.0          | 33.7   | 58.0   | 68.4   |
| RetinaNet   | Resnet101-FPN     | 32.7        | 52.0         | 55.3          | 34.4   | 59.8   | 70.6   |
| OptNet      | Resnet50-FPN      | 32.2        | 51.7         | 55.1          | 35.1   | 59.2   | 69.4   |
| OptNet      | Resnet101-FPN     | 33.2        | 52.9         | 56.3          | 35.8   | 60.5   | 71.4   |

TABLE 3. Effectiveness of each module on COCO val.

| Attention Feature | Feature Align | CS | AP | AP_{50} | AP_{75} | AP_{S} | AP_{M} | AP_{L} |
|------------------|---------------|----|----|---------|---------|--------|--------|--------|
| -                | -             | -  | 36.0| 56.1    | 38.1    | 20.3   | 40.0   | 46.6   |
| -                | ✓             | ✓  | 36.5| 56.1    | 39.0    | 20.2   | 40.4   | 48.0   |
| ✓                | -             | -  | 36.8| 56.5    | 39.7    | 20.6   | 40.5   | 47.8   |
| ✓                | ✓             | ✓  | 36.8| 56.9    | 39.4    | 20.3   | 40.7   | 48.1   |
| ✓                | ✓             | ✓  | 37.1| 57.6    | 39.5    | 21.2   | 40.9   | 48.4   |

This enhances the discriminability of the features. Then, the obtained features and results of the first prediction are fused for the second classification and regression prediction of the object. In the experiment, this operation promotes the detection performance because the first prediction results are learned with the real object information. In addition, the influence of outliers on the detector is eliminated to a certain extent.

**E. LOSS FUNCTION**

OptNet’s loss function is composed of classification and localization loss in the first stage, and classification and localization loss in the second stage. Both stages use the same labels to achieve supervised learning of the task. OptNet uses focal loss and smooth L1 loss as the loss functions for classification and localization, respectively.

\[
L_{cls}^t = \frac{1}{N_{pos}} \left( \sum_{i \in Pos} FL(p_i^t, p_i^*) + \sum_{i \in Neg} FL(p_i^t, p_i^*) \right) \quad (2)
\]

\[
L_{loc}^t = \frac{1}{N_{pos}} \left( \sum_{i \in Pos} \sum_{j \in Pos} smooth_{L1}(r_i^t, r_j^*) \right) \quad (3)
\]

where \( L_{cls}^t \) and \( L_{loc}^t \) represent the classification and localization loss at stage \( t \), respectively. \( i \) represents the index of the anchor box in a minibatch, and \( p_i^t \) and \( p_i^* \) represent the probabilities that the anchor box is the object and its corresponding true label, respectively. \( r_i^t \) and \( r_i^* \) represent the predicted offsets of the anchor box and its corresponding true offset, respectively. \( N_{pos} \) is the number of positive samples in a minibatch. In the experiment, multiple classifications and regressions optimize the training of the network and improve the precision of detection.

**IV. EXPERIMENTS**

**A. TRAINING AND TESTING DETAILS**

OptNet is based on the detection framework of RetinaNet, and all experiments are conducted on MMDetection. In the training stage, OptNet generates three anchor boxes of different scales at each spatial location of the feature map. To obtain the
positive and negative samples required for training, OptNet first calculates the IoU of these anchor boxes and all ground true boxes and then divides all the anchor boxes into positive and negative samples in a certain way. This specific method is the same as RetinaNet. All experiments are conducted on two GPUs, and the linear rule adjusts the learning rate of the network. To better confirm the effectiveness of the algorithm, the batch size of ResNet50 is set to 10 in this work (five images per GPU), and the learning rate is set as 0.01. For the backbone network ResNet101, the batch size is set to 6 (three images per GPU), the learning rate is set to 0.005, and all other parameters follow the original MMDetection configuration. In the test phase, the detector obtains a large number of prediction boxes, which have different degrees of redundancy. To obtain a more accurate object prediction box, the non-maximum suppression (NMS) algorithm is used to remove the repeated box. OptNet only uses the classification and regression results of the second prediction as the final result of the detector.

### B. EXPERIMENTAL RESULTS

Different backbone networks are used on OptNet. RetinaNet and other methods are reimplemented on MMDetection for a fair comparison. Table 1 lists the detection results of OptNet and other detection algorithms on the MS COCO dataset. In the same experimental environment, the AP of OptNet on different backbone networks is 0.9%~1.1% higher than that of RetinaNet, which demonstrates that OptNet can effectively improve the performance of classification and localization. To better display the object detection performance of

---

**TABLE 4.** Detection performance of two results of network prediction on MS COCO dataset.

| Setting          | Backbone | AP  | AP<sub>50</sub> | AP<sub>75</sub> | AP<sub>S</sub> | AP<sub>M</sub> | AP<sub>L</sub> |
|------------------|----------|-----|-----------------|-----------------|---------------|---------------|---------------|
| OptNet-first      | ResNet50 | 36.8| 57.1            | 39.5            | 21.1          | 40.8          | 48.1          |
| OptNet-first      | ResNet101| 38.7| 58.9            | 41.6            | 22.1          | 43.0          | 50.5          |
| OptNet-second     | ResNet50 | 37.1| 57.6            | 39.5            | 21.2          | 40.9          | 48.4          |
| OptNet-second     | ResNet101| 39.0| 59.0            | 41.9            | 22.2          | 43.2          | 51.1          |

**TABLE 5.** Results of multiple attention feature module design schemes on MS COCO dataset.

| Setting | AP  | AP<sub>50</sub> | AP<sub>75</sub> | AP<sub>S</sub> | AP<sub>M</sub> | AP<sub>L</sub> |
|---------|-----|-----------------|-----------------|---------------|---------------|---------------|
| a       | 36.6| 56.5            | 39.2            | 20.5          | 40.1          | 47.5          |
| b       | 36.7| 56.4            | 39.3            | 20.6          | 40.9          | 47.4          |
| c       | 37.0| 57.0            | 39.7            | 21.1          | 40.7          | 48.4          |
| d       | **37.1**| **57.6**| **39.5**| **21.2**| **40.9**| **48.4**|
FIGURE 7. Visualization results of OptNet and RetinaNet. Blue box is prediction result of detector, yellow box is misdetection result, and red box is missed detection.

OptNet for different scales, the average precision of small objects, medium objects and large objects are denoted by $AP_S$, $AP_M$ and $AP_L$, respectively. For $AP_S$, OptNet increases the detection precision by $0.8\%$-$1.2\%$ because the low-level feature map obtains better small-object features. For $AP_M$ and $AP_L$, OptNet also has remarkable improvements. To better illustrate the effectiveness of our method, OptNet is compared with two-stage detection models. With the same backbone, the AP of OptNet surpasses that of Faster R-CNN by $0.6\%$ and $1.3\%$. These improvements are caused by the discriminative object uniqueness information and accurate localization in OptNet. In addition, we compare the recall rates of OptNet and RetinaNet. $AR_S$, $AR_L$ and $AR_M$ represent the average recall rates of small objects, medium objects and large objects. As shown in Table 2, our method outperforms RetinaNet for all recall rate indicators. For $AR_{max}=10$, OptNet increases the recall rate by $0.9\%$ and $1.0\%$ on different backbone networks. The performance for $AR_{max}=100$ is improved...
by 1% and 1.1%. Thus, our method can remarkably improve the recall rate of the overall prediction box of the detector. In addition, OptNet has different degrees of improvement in $\text{AR}_S$, $\text{AR}_M$ and $\text{AR}_L$, especially for the recall rate of small objects. Thus, our method can better detect multi-scale objects.

C. ABLATION EXPERIMENTS

OptNet uses three modules to enhance the detection performance. The ablation experiments are conducted on the MS COCO Val dataset using Resnet50 as the backbone network. Table 3 shows the impact of each module on the performance. All three modules improve the detection performance. Attention feature model (AFM) improves the performance by 0.6%, which is not very remarkable, but it improves all of the detectors on the whole. The average precisions of small, medium and large objects increase by 1%, 0.5%, and 0.4%, respectively. This indicates that AFM can extract better object features, especially small object features.

Feature Align uses the offset predicted by the network to obtain more accurate object features. This module can remarkably improve the detection performance of small objects the most. Since the features of medium objects and large objects are obvious, feature alignment has little impact on them.

CS improves the detection performance of small, medium and large objects by 0.6%, 0.4% and 0.6%, respectively. This indicates that consistency supervision can build a better feature pyramid. In summary, the three modules proposed in this paper are effective. The AP of the OptNet detector is 1.1% higher than that of the original algorithm. In addition, OptNet improves the detection performance for objects of different scales, especially for large objects (improved by 1.8%).

As shown in Figure 3, OptNet implements two predictions for object category and localization. Table 4 compares the performance of the two predicted results on the MS COCO dataset. The performance of the second prediction is 0.3% higher than that of the first prediction with regard to the average precision. The detection performance improvements are remarkable for large objects, as shown by the second prediction result.

To explore an excellent feature pyramid network, Figure 6 shows four design schemes of attention feature modules. In Figure 6 (a), feature map $C_3$ is convolved with two $3 \times 3$ convolutions. Note that BN and ReLU follow each convolution. Because feature extraction is too simple, its effect is unsatisfactory. In Figure 6 (b), first, a $1 \times 1$ reduces the dimensionality of features, and a $3 \times 3$ convolution extracts object features. Then, high-level features are fused with low-level features. Finally, the fusion feature is optimized by $3 \times 3$ deformable convolution. Neither the feature loss nor the network can learn the location shift adaptively, and the scheme is also unsatisfactory. In Figure 6 (c), three $3 \times 3$ convolutions extract the features of three different scales, and then the high-level features and low-level features are fused. This design scheme exhibits good performance. Figure 6 (d) shows the design scheme in this paper, which has satisfactory performance. Table 5 compares the schemes. The experimental results demonstrate that satisfactory performance can be obtained by using several $3 \times 3$ convolutions and embedding attention modules.

To further illustrate the effectiveness of OptNet, Figure 7 shows the results of RetinaNet and OptNet. The experimental images cover multiple scenes, and each image has many objects and occlusions. The first column in Figure 7 shows the detection results of the same image using different detectors, and the second column shows the localization results. OptNet and RetinaNet identify all objects in the image, so both detectors have good detection performance. However, the localization performance of OptNet is better than that of RetinaNet, which indicates that the optimized regression can improve the localization precision. In the third and fourth lines of Figure 7, OptNet generally obtains higher classification scores than RetinaNet and detects more small objects, but OptNet identifies the stem flower in the upper right corner as a vase, which indicates that OptNet can understand the scene. At the same time, OptNet and RetinaNet produce a large number of prediction boxes, and these boxes contain multiple objects. In the fifth and sixth lines, OptNet and RetinaNet can accurately identify the objects, and the classification score of each object is high. However, RetinaNet misses a large number of small objects and severely occluded objects, while OptNet obtains richer object information. Through the above analysis, the detection performance of OptNet is better than that of RetinaNet in complex scenes. Furthermore, OptNet has more accurate localization and higher confidence and can detect more small objects and occluded objects.

V. CONCLUSION AND FURTHER WORKS

The extraction of discriminative object features and accurate localization are highly important in object detection. This paper introduced and analyzed feature extraction in object detection. Attention feature and feature alignment modules were designed to extract more discriminative object features. At the same time, a consistency supervision module was proposed to supervise learning on the features of the feature pyramid and optimize the results of the first prediction. Experiments showed that OptNet is effective and can detect more small objects even in complex scenes. However, the method proposed in this paper increases the parameters of the model, so its detection speed is slower than that of RetinaNet. In future work, we will improve the attention mechanisms and embed them into detection networks to achieve better object detection. In addition, we will also try to compress the network to reduce the number of model parameters and accelerate the speed.

REFERENCES

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.
JIALE XIONG received the bachelor’s degree from Nanchang Hangkong University, Jiangxi, China, where he is currently pursuing the master’s degree. His research interests include computer vision, image processing, and pattern recognition.

JIEXIAN ZENG received the master’s degree in engineering from Northwestern Polytechnical University, in 1997. He is currently a Professor with Nanchang Hangkong University. He has published over 120 papers. His research interests include image processing, pattern recognition, and computer vision research.

XIANG FU received the Ph.D. degree from Xidian University, in 2008. He is currently an Associate Professor with Nanchang Hangkong University. His research interests include computer vision, image processing, and pattern recognition.

LU LENG (Member, IEEE) received the Ph.D. degree from Southwest Jiaotong University, Chengdu, China, in 2012. He was a Postdoctoral Researcher with Yonsei University, Seoul, South Korea, and Nanjing University of Aeronautics and Astronautics, Nanjing, China. He was also a Visiting Scholar with West Virginia University, Morgantown, WV, USA, and Yonsei University. He is currently an Associate Professor with Nanchang Hangkong University. He has published more than 100 international journals and conference papers. He has been granted several scholarships and funding projects in his academic research. His research interests include computer vision, biometric template protection, and biometric recognition.

Dr. Leng is a member of the Association for Computing Machinery (ACM), China Society of Image and Graphics (CSIG), and China Computer Federation (CCF). He is a reviewer of several international journals and conferences.