

You Should Look at All Objects

Zhenchao Jin¹ , Dongdong Yu², Luchuan Song³, Zehuan Yuan², and Lequan Yu¹ *

¹ The University of Hong Kong

blwx96@connect.hku.hk, lqyu@hku.hk

² Bytedance

{yudongdong, yuanzehuan}@bytedance.com

³ University of Rochester

lsong11@ur.rochester.edu

Abstract. Feature pyramid network (FPN) is one of the key components for object detectors. However, there is a long-standing puzzle for researchers that the detection performance of large-scale objects are usually suppressed after introducing FPN. To this end, this paper first revisits FPN in the detection framework and reveals the nature of the success of FPN from the perspective of optimization. Then, we point out that the degraded performance of large-scale objects is due to the arising of improper back-propagation paths after integrating FPN. It makes each level of the backbone network only has the ability to look at the objects within a certain scale range. Based on these analysis, two feasible strategies are proposed to enable each level of the backbone to look at all objects in the FPN-based detection frameworks. Specifically, one is to introduce auxiliary objective functions to make each backbone level directly receive the back-propagation signals of various-scale objects during training. The other is to construct the feature pyramid in a more reasonable way to avoid the irrational back-propagation paths. Extensive experiments on the COCO benchmark validate the soundness of our analysis and the effectiveness of our methods. Without bells and whistles, we demonstrate that our method achieves solid improvements (more than 2%) on various detection frameworks: one-stage, two-stage, anchor-based, anchor-free and transformer-based detectors ⁴.

Keywords: Object Detection, Feature Pyramid Network

1 Introduction

Along with the advances in deep neural networks, recent years have seen remarkable progress in object detection, which aims at detecting objects of predefined categories. A common belief for the success of the state-of-the-art detectors [2, 9, 14, 39, 45] is the use of feature pyramid network (FPN) [21]. Despite impressive, there is an unexpected phenomenon after introducing FPN that the

* Corresponding author.

⁴ Our code will be available at <https://github.com/CharlesPikachu/YSLAO>.

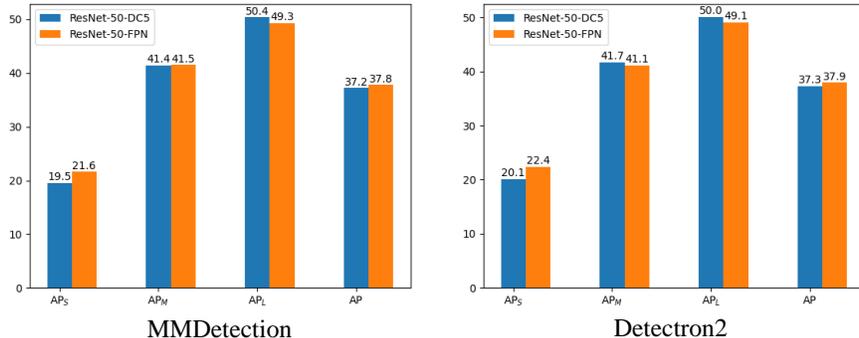


Fig. 1. Comparing the detection performance between ResNet-50-DC5 and ResNet-50-FPN based on MMDetection [4] and Detectron2 [42] toolboxes. The adopted detector is Faster R-CNN [31]. The detectors are trained on COCO 2017 train set and evaluated on COCO 2017 validation set [23].

overall detection performance improvement is built upon the *increased* Average Precision of small objects (AP_S) and the *decreased* Average Precision of large objects (AP_L). For instance, the experiments based on MMDetection [4] and Detectron2 [42] in Figure 1 demonstrate this phenomenon. When we leverage the detection toolbox MMDetection, we can observe that AP_S increases from 19.5% to 21.6% while AP_L decreases from 50.4% to 49.3% after integrating FPN. The consistent trend can also be observed in Detectron2.

Prior to this study, there are mainly two assumptions on why the introduction of FPN works. The first is that the use of FPN helps obtain better representations by fusing multiple low-level and high-level feature maps [7, 17, 21, 24, 29]. The second is that each pyramid level can be responsible for detecting objects within a certain scale range, *i.e.*, divide-and-conquer [5]. Obviously, both assumptions should lead to the same conclusion that the increase in AP is due to the co-increase in AP_S, AP_M and AP_L. However, the unexpected drops of AP_L in Figure 1 indicates that there are other key differences between FPN-free and FPN-based detection frameworks, while few studies have taken note of this. In this paper, we propose to investigate FPN from the perspective of optimization. Our assumption is that *except for the multi-scale feature fusion and divide-and-conquer, the back-propagation paths altered by FPN will also directly influence the performance of the detection frameworks.*

We start from explaining why FPN can benefit the detection framework by changing the back-propagation paths. Then, we point out that the back-propagation paths altered by the existing FPN paradigm will make each backbone stage only have the ability to see the objects within a certain scale range (*i.e.*, extracting features that are only fit to certain scale range objects), which is the cause of the inconsistent changes in AP_S, AP_M and AP_L in Figure 1. Accordingly, the key insight to achieve the consistent improvements in AP of the objects with various scale ranges is to enable each backbone stage to see all objects during training. Based on this principle, we propose to expand and amend the existing back-propagation paths in FPN-based detection frameworks.

Our approach of expanding the back-propagation paths is to introduce auxiliary objective functions so that both the original signals and the extra signals can jointly supervise the learning of the corresponding backbone levels. The key technique to the success of this approach is to introduce the uncertainty [15, 16] to better balance the various back-propagation signals. The strategy of amending the back-propagation paths is to build the feature pyramid in a more effective way and thereby, all levels of the backbone network can receive the sufficient signals. The key technique of this approach is the feature grouping module used to promise the space compactness of homogeneous representations.

In a nutshell, the contributions of this paper are:

- To the best of our knowledge, it is the first work to reveal the nature of the success of FPN from the perspective of optimization. Further, we provide new insight in explaining why the introduction of the traditional FPN would suppress the performance of large-scale objects from this perspective.
- We propose to introduce auxiliary objective functions guided by uncertainty to mitigate the inconsistent changes in AP_S , AP_M and AP_L . Since there are no additional computational overhead during testing in the strategy, the inference speed of the detectors can be preserved from decreasing.
- We propose a novel feature pyramid generalization paradigm. The key idea is to make the back-propagation signals of various-scale objects can directly pass to each level of the backbone network. We further design a cascade structure to achieve more robust Average Precision (AP) improvements.
- The extensive experiments on COCO benchmark validate the soundness of our principle and the effectiveness of our solutions. Without bells and whistles, our method boosts the detection performance by more than 2% AP on various frameworks: one-stage, two-stage, anchor-based, anchor-free and transformer-based detectors.

2 Related Works

Object Detection. Recent years have witnessed remarkable improvements in object detection [2, 14, 29, 37, 39]. In general, there are two leading paradigms in this area, *i.e.*, one-stage and two-stage frameworks. Two-stage pipeline is first introduced by R-CNN [8], where a set of region proposals are yielded in the first stage, and then the second stage classifies and refines the proposals. The next milestone of two-stage detector is the emergence of Faster R-CNN [31], which aims to improve the efficiency of two-stage methods and allow the detectors to be trained end-to-end. After that, plenty of algorithms have been proposed to further boost its performance, including applying multi-scale training and testing [35, 36], redesigning and reforming architecture [2, 3, 9, 41, 50], introducing relation and attention mechanism [13, 26, 32], improving the training strategy and loss function [12, 20, 28, 30, 33], adopting more reasonable post-processing algorithms [1, 11, 25, 40]. Different from the two-stage approaches, one-stage detectors directly predict the object category and location based on the predefined

anchors. They are simpler and faster than two-stage methods but have trailed the detection performance until the emergence of RetinaNet [22]. Thereafter, lots of works [5, 6, 19, 39] are presented to boost the detection performance of one-stage detectors and at present, one-stage methods can achieve very close performance with two-stage frameworks at a faster inference speed.

Feature Pyramids. Feature pyramids have dominated modern detectors for several years. Recent researches on feature pyramids can be roughly categorized into three gatherings: top-down or bottom-up networks [21, 24, 29, 34], attention based methods [14, 17, 44, 47], and neural architecture search based approaches [7, 38]. Specifically, feature pyramid network (FPN) [21] is one of the most classical paradigms to build a feature pyramid, which designs a top-down architecture with lateral connections to make each pyramid level carry the high-level semantic information. After that, several works [7, 14, 17, 24, 29, 47] follow FPN and make attempts to obtain more effective representations by improving the strategy of multi-scale feature fusion. PANet [24] proposes to leverage bottom-up architecture to shorten the information interaction path between shallow layers and topmost features. SAFNet [14] aims to suppress the redundant information at all pyramid scales by introducing attention mechanism. Nas-fpn [7] proposes to construct the feature pyramids by neural architecture search. However, the starting point of the above methods is that FPN can bring two benefits, *i.e.*, leveraging multi-scale feature fusion to obtain more effective representations [7, 17, 21, 24, 29] and adopting divide-and-conquer to reduce the learning difficulty [5]. And it fails to explain why introducing FPN will suppress the performance of large-scale objects. Motivated by this, we propose to revisit FPN from the perspective of optimization, which successfully explains the anomalous phenomenon in Figure 1. From this novel starting point, we further propose to mitigate the inconsistent changes in AP_S , AP_M and AP_L by expanding or amending the back-propagation paths in FPN-based detection frameworks. And it is the main difference between our approaches and previous works.

3 Revisit FPN

3.1 Backbone Network

In object detection, the backbone network \mathcal{F}_B is used to extract the basic features \mathcal{C} from the input image I . For the convenience of presentation, we assume that the adopted backbone network is ResNet [10]. It generally consists of one basic feature extractor and plenty of residual blocks, where the residual blocks can be grouped into four stages according to the resolutions of the output feature maps. Specifically, \mathcal{C} is calculated as follow,

$$\begin{aligned} C_1 &= f_{s_0}(I), \\ C_i &= f_{s_{i-1}}(C_{i-1}), \quad 2 \leq i \leq 5, \end{aligned} \tag{1}$$

where \mathcal{C} consists of $\{C_2, C_3, C_4, C_5\}$ and \mathcal{F}_B consists of $\{f_{s_0}, f_{s_1}, f_{s_2}, f_{s_3}, f_{s_4}\}$.

3.2 FPN-free Detection Framework

For FPN-free detectors, the network usually leverages C_5 to perform the classification and regression of the objects as follow,

$$\begin{aligned} O_{cls} &= f_{cls}(f_{pre}(C_5)), \\ O_{reg} &= f_{reg}(f_{pre}(C_5)), \end{aligned} \quad (2)$$

where f_{pre} is introduced to unify various operations between C_5 and the output results, *e.g.*, the region proposal network [31]. O_{cls} and O_{reg} are the predicted category information and location information of the objects, respectively. f_{cls} and f_{reg} are a 1×1 convolution layer, respectively. During training, the classification and regression loss are calculated as follow,

$$L = L_{cls}(O_{cls}, GT_{cls}) + \lambda L_{reg}(O_{reg}, GT_{reg}), \quad (3)$$

where the adopted objective functions L_{cls} and L_{reg} depend on the utilized detection framework. GT_{cls} and GT_{reg} are the ground-truth classification and regression information, respectively. λ is a hyper-parameter used to balance the classification and regression losses.

3.3 FPN-based Detection Framework

For FPN-based detectors, \mathcal{C} is first used to build the feature pyramid as follow,

$$\begin{aligned} C'_5 &= flat_5(C_5), \\ C'_4 &= flat_4(C_4) + UP_{2 \times}(C'_5), \\ C'_3 &= flat_3(C_3) + UP_{2 \times}(C'_4), \\ C'_2 &= flat_2(C_2) + UP_{2 \times}(C'_3), \\ P_l &= f_{smo_l}(C'_l), \quad 2 \leq l \leq 5, \end{aligned} \quad (4)$$

where $\mathcal{P} = \{P_2, P_3, P_4, P_5\}$ is the constructed feature pyramid. $UP_{2 \times}$ denotes for the upsampling with the scale factor of 2. $flat_i, 2 \leq i \leq 5$ is the lateral connections implemented by a 1×1 convolution layer, respectively, which is used to change the number of the channels of \mathcal{C} . $f_{smo_l}, 2 \leq l \leq 5$ is a linear function and is usually implemented by a 3×3 convolution layer. Without loss of generality, Eq.(4) can be rewritten as follow,

$$P_l = \sum_{i=l}^5 w_i \cdot C_i, \quad 2 \leq l \leq 5, \quad (5)$$

where w_i is the final weights for the correspond level after polynomial expansions [17]. Then, the network uses \mathcal{P} to predict the classification and regression information of the objects assigned to each pyramid level l as follow,

$$\begin{aligned} O_{cls,l} &= f_{cls,l}(f_{pre,l}(P_l)), \\ O_{reg,l} &= f_{reg,l}(f_{pre,l}(P_l)). \end{aligned} \quad (6)$$

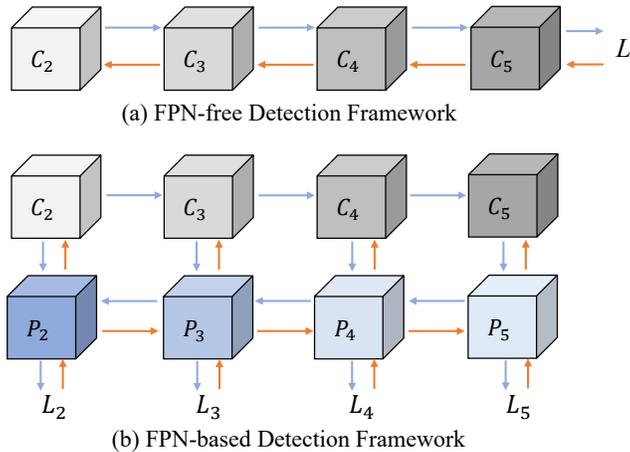


Fig. 2. Comparing the back-propagation paths between FPN-free detection framework and FPN-based detection framework. The blue arrows represent for forward and the orange arrows denote for back propagation. Note that, only the most significant back-propagation signals to each backbone level will be marked.

The objects assignment rule is to make the low-resolution pyramid features (*e.g.*, P_5) be responsible for predicting the large-scale objects, while the high-resolution pyramid features (*e.g.*, P_2) are utilized to predict the small-scale objects. During network optimization, the losses at each pyramid level l are calculated as follow,

$$L_l = L_{cls}(O_{cls,l}, GT_{cls,l}) + \lambda L_{reg}(O_{reg,l}, GT_{reg,l}). \quad (7)$$

3.4 Analysis of FPN

From Section 3.2 and 3.3, we can observe that introducing FPN can alter the back-propagation paths between the objective functions and the backbone network. Figure 2 shows the detailed differences between the FPN-free and FPN-based detection pipeline. In the FPN-free detection pipeline, only the backbone feature C_5 is directly under the supervision of the objective functions. Since there exists the vanishing gradient problem in deep neural networks, the shallow layers (*i.e.*, $\{f_{s_0}, f_{s_1}, f_{s_2}, f_{s_3}\}$) of the backbone network will be difficult to receive effective supervision by the backward propagation. While in the FPN-based detection framework, we can observe that all the backbone features are directly under the supervision of the objective functions. Since this strategy avoids the vanishing gradient problem for the shallow layers, each level of the backbone network can receive more supervision to train its own parameters. We believe that it is the key reason why FPN-based detectors outperform FPN-free detectors from the perspective of optimization.

To further demonstrate the principle above, we conduct the empirical study and show the experimental results in Figure 3. FPN-Aux and DC5-Aux denote

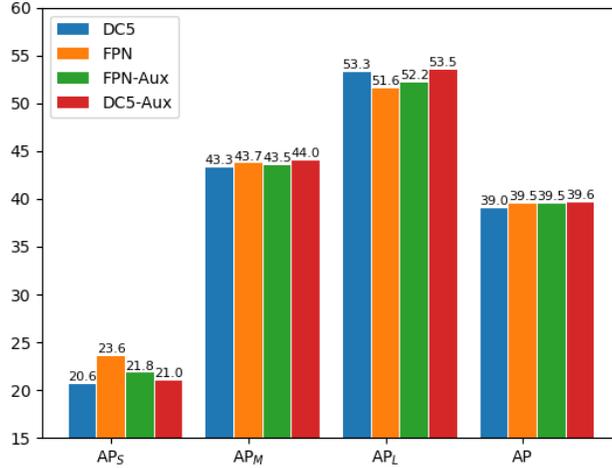


Fig. 3. The detection performance with different settings. The adopted backbone network is ResNet-101 and the utilized detector is Faster R-CNN. The models are trained on COCO 2017 train set and evaluated on COCO 2017 validation set [23].

for introducing the auxiliary losses in the shallow layers of the backbone networks [40, 48]. Specifically, given \mathcal{C} , we first have

$$\begin{aligned}\hat{O}_{cls,i} &= \hat{f}_{cls,i}(\hat{f}_{pre,i}(C_i)), \quad 2 \leq i \leq 4, \\ \hat{O}_{reg,i} &= \hat{f}_{reg,i}(\hat{f}_{pre,i}(C_i)), \quad 2 \leq i \leq 4.\end{aligned}\quad (8)$$

For the two-stage detectors [31], to avoid double calculation of the proposals, we will utilize the proposals calculated in Eq.(2) or Eq.(6) to extract the ROIs. Then, the auxiliary losses can be obtained as follow,

$$\hat{L}_i = L_{cls}(\hat{O}_{cls,i}, GT_{cls}) + \lambda L_{reg}(\hat{O}_{reg,i}, GT_{reg}). \quad (9)$$

And the final loss of the detection framework is the summation of the auxiliary losses and the original losses. Since the auxiliary losses can be used to directly supervise the learning of the shallow layers of the backbone network, if our assumption is correct, introducing auxiliary losses should own a similar function to integrating FPN from the perspective of optimization. In Figure 3, it is observed that the auxiliary losses can boost the detection performance of FPN-free detector (from 39.0% to 39.6%) and obtain a comparable AP result to FPN-based detector (39.6% *v.s.* 39.5%). However, the introduction of auxiliary losses seems useless to FPN-based detector (from 39.5% to 39.5%). This result validates our assumption that from the perspective of optimization, the nature of the success of FPN is the shorten back-propagation distance between the objective losses and the shallow layers of the backbone network.

Now, the question is why the introduction of FPN will suppress the detection performance of large-scale objects. As illustrated in Figure 2, P_2 is a linear combination of $\{C_2, C_3, C_4, C_5\}$, therefore, L_2 can directly supervise the learning

of all the backbone stages. With the similar principle, L_3 , L_4 and L_5 have the ability to directly constrain $\{C_3, C_4, C_5\}$, $\{C_4, C_5\}$ and C_5 , respectively. However, as mentioned above, L_2 is only used to make the corresponding backbone levels focus on the objects within a small scale range. Thus, the learned feature C_2 only has the ability to detect the small-scale objects well by back propagation. Meanwhile, the backbone network also needs to utilize Eq.(1) to calculate $\{C_3, C_4, C_5\}$ with C_2 as input. Obviously, it is insufficient for f_{s_2} to extract rich semantic features of the larger objects from C_2 . Worse still, the adverse effects will be further accumulated when leveraging f_{s_3} and f_{s_4} to calculate C_4 and C_5 . As a result, the semantic information carried by C_5 is somehow ineffective for predicting the large-scale objects. And it is why there is always an unexpected phenomenon after introducing FPN that the overall detection performance improvement is built upon the increased AP_S and the decreased AP_L .

The empirical study in Figure 3 also validate our assumption. In detail, we can observe that after applying the auxiliary losses to the FPN-based detection frameworks, the performance improvements among $\{AP_S, AP_M, AP_L\}$ tend to be consistent with FPN-free detectors w/ the auxiliary losses. The result shows that as the auxiliary losses can help the shallow layers of the backbone learn the features to detect various-scale objects, C_5 no longer suffers from the ineffective features for predicting large-scale objects as only integrating FPN into the detection framework. In other words, the lack of effective semantic information of large-scale objects in C_5 is the key reason for the decrease of AP_L . And the problem is derived from the inability of $f_{s_i}, 1 \leq i \leq 3$ to look at various-scale objects during training.

4 Methodology

Motivated by the finding that the inconsistent changes in $\{AP_S, AP_M, AP_L\}$ is caused by the inability of $f_{s_i}, 1 \leq i \leq 3$ to see all objects during training, we propose to make the backbone stages look at various-scale objects by expanding or amending the back-propagation paths in FPN-based detection frameworks to address the decreased AP_L problem above. Specifically, we propose two strategies, *i.e.*, introducing auxiliary objective functions and building the feature pyramid in a more reasonable manner in this section.

4.1 Auxiliary Losses

As mentioned in Section 3.4, introducing auxiliary losses can help $f_{s_i}, 1 \leq i \leq 3$ own the ability to see all objects. However, the simple summation of the losses may be insufficient. In order to introduce auxiliary losses more rationally, we propose to leverage the uncertainty [15, 16, 43] to better balance the various-type loss signals. Specifically, we incorporate the uncertainty into each classification and regression auxiliary loss as follow,

$$\mathcal{L}(p, gt) = e^{-\alpha} \hat{\mathcal{L}}(p, gt) + \tau\alpha, \quad (10)$$

where p is the predicted result and gt is the corresponding ground truth. $\hat{\mathcal{L}}$ denotes for the loss function, *e.g.*, L_{reg} and L_{cls} . τ is a hyper-parameter used to avoid generating high uncertainty α . α is generated as follow,

$$\alpha = ReLU(w \cdot x + b), \quad (11)$$

where x is the feature map also utilized to predict p . w and b are the learnable parameters. $ReLU$ is used to promise $\alpha \geq 0$.

4.2 Feature Pyramid Generation Paradigm

Building the feature pyramid in a more reasonable way is also an effective method to achieve consistent improvements in $\{AP_S, AP_M, AP_L\}$. As analysed in Section 3.4, the problem in the process of constructing traditional FPN is caused by Eq.(4). Specifically, P_l should contain the feature maps from all backbone levels so that L_l can help each backbone level see the objects inputted to L_l . Accordingly, the summation of $L_l, 2 \leq l \leq 5$ can make each backbone level own the ability to look at all objects.

Feature Grouping. To select effective feature maps from $\mathcal{C}' = \{C'_2, C'_3, C'_4, C'_5\}$ for the objects assigned to the corresponding pyramid level, we first perform channel swapping on \mathcal{C}' as follow,

$$X_k = R^{zhw}(M_k \otimes R^{zn}(C'_k)), \quad 2 \leq k \leq 5, \quad (12)$$

where \otimes denotes for matrix multiplication. R^{zn} reshapes C'_k into the size of $Z \times HW$ and R^{zhw} reshapes the input tensor into the size of $Z \times H \times W$, where Z is the number of the channels and $H \times W$ is the resolution of the feature map. M_k is a matrix of size $Z \times Z$ used to achieve channel swapping. In practice, M_k is generated as follow,

$$M_k = G_k(C'_k), \quad (13)$$

where the structure of G_k is shown in Figure 4. We expect M_k to own the ability to make the homogeneous feature maps become compact along the channel dimension.

Then, X_k is divided into quarters along the channel dimension,

$$X_k = \{X_{k,2}, X_{k,3}, X_{k,4}, X_{k,5}\}, \quad (14)$$

where we assume that $X_{k,l}, 2 \leq l \leq 5$ only carries the effective semantic information of the objects assigned to the pyramid level l . After that, we have

$$P'_l = X_{2,l} \oplus X_{3,l} \oplus X_{4,l} \oplus X_{5,l}, \quad 2 \leq l \leq 5, \quad (15)$$

where \oplus denotes for the concatenation operation. Finally, the feature pyramid is constructed as follow,

$$P_l = f_{smoi}(P'_l), \quad 2 \leq l \leq 5. \quad (16)$$

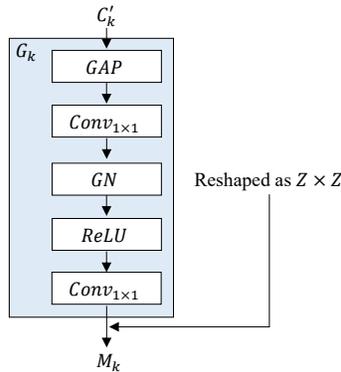


Fig. 4. An illustration of the structure of G_k . $Conv_{1 \times 1}$ denotes for a 1×1 convolution layer. GN means group normalization and GAP is the global average pooling operation.

Cascade Structure. To better promote the space compactness, we propose to employ a cascade structure to conduct feature grouping in a coarse-to-fine manner. Specifically, at the second stage, P'_l will first be taken as the input of the feature grouping module and thereby we can obtain \hat{P}'_l . Then, we have

$$P''_l = f_w(P'_l) \cdot P'_l + f_w(\hat{P}'_l) \cdot \hat{P}'_l, \quad 2 \leq l \leq 5, \quad (17)$$

where f_w is a non-linear function used to generate the feature fusion weights. In our implementation, f_w consists of two convolution blocks (a block consists of a convolution, a normalization and an activation layer). Finally, Eq.(16) will be conducted to obtain the feature pyramid with the input P''_l . The same can be done for the cases when the number of the stages is greater than 2.

5 Experiments

Dataset. Our approaches are evaluated on the challenging MS COCO benchmark [23], which contains $\sim 118k$ images for training (*train-2017*), $5k$ images for validation (*val-2017*) and $\sim 20k$ images with no disclosed annotations for testing (*test-dev*). By default, the detection frameworks in this section are trained on *train-2017* set and evaluated on *val-2017* set.

Implementation Details. Our methods are implemented with MMDetection [4]. We train our detection frameworks on 8 NVIDIA Tesla V100 GPUs with a 32 GB memory per-card. Following previous works [21, 22, 31], we initialize the backbone networks using the weights pre-trained on ImageNet [18] and randomly initialize the weights of the newly added modules. The input images are resized to keep their shorter side being 800 and their longer side less or equal to 1,333. The optimizer is stochastic gradient descent (SGD) with momentum of 0.9, weight decay of 0.0001, and batch size of 16 (*i.e.*, 2 images per GPU). By default, the models are trained for 12 epochs ($1 \times$ schedule), and we set the initial learning rate as 0.02 and decay it by 0.1 at epoch 9 and 11, respectively. We adopt random

Table 1. Ablation study on auxiliary losses. FPS is evaluated on a single Titan Xp.

Framework	Backbone	Auxiliary	Uncertainty	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	FPS
<i>One-stage</i>										
RetinaNet	ResNet-101			38.5	57.6	41.0	21.7	42.8	50.4	15.0
RetinaNet	ResNet-101	✓		38.7	57.7	41.2	21.2	43.0	51.2	15.0
RetinaNet	ResNet-101	✓	✓	40.1	61.4	43.7	23.3	44.4	52.4	15.0
<i>Two-stage</i>										
Faster R-CNN	ResNet-101			39.5	60.4	42.9	23.6	43.7	51.6	15.6
Faster R-CNN	ResNet-101	✓		39.5	60.0	43.3	21.8	43.5	52.2	15.6
Faster R-CNN	ResNet-101	✓	✓	40.9	62.0	44.8	24.2	45.3	53.3	15.6

Table 2. Ablation study on the feature pyramid generation paradigm.

Framework	Backbone	Feature Grouping	Cascade Times	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	FPS
<i>One-stage</i>										
RetinaNet	ResNet-101			38.5	57.6	41.0	21.7	42.8	50.4	15.0
RetinaNet	ResNet-101	✓	1×	40.2	60.1	42.6	23.3	44.5	52.7	12.2
RetinaNet	ResNet-101	✓	2×	40.8	60.5	43.6	24.0	44.8	54.4	11.7
RetinaNet	ResNet-101	✓	3×	41.2	60.8	43.8	24.1	45.2	55.2	11.1
<i>Two-stage</i>										
Faster R-CNN	ResNet-101			39.5	60.4	42.9	23.6	43.7	51.6	15.6
Faster R-CNN	ResNet-101	✓	1×	40.6	61.9	44.5	24.2	45.0	52.7	12.0
Faster R-CNN	ResNet-101	✓	2×	41.7	62.7	45.4	24.8	45.9	53.7	11.4
Faster R-CNN	ResNet-101	✓	3×	42.2	63.0	45.8	25.5	46.1	55.8	10.9

horizontal flip as the data augmentation. Other unmentioned hyper-parameters follow the settings in MMDetection.

In the inference phase, the input image is first resized in the same way as the training phase and then we forward it through the whole network to output the predicted bounding boxes with the category probability distribution. After that, we leverage a score 0.05 to preliminary filter out background bounding boxes and then output the top 1,000 detections per pyramid level. Finally, the non-maximum suppression (NMS) is applied with the IoU threshold 0.5 per class to output the final top 100 confident detections per image.

Evaluation Metrics. The results are evaluated with standard COCO-style metrics, including AP (averaged over IoU thresholds), AP₅₀ (AP for IoU threshold 50%), AP₇₅ (AP for IoU threshold 75%), AP_S (AP on objects of small scales), AP_M (AP on objects of medium scales) and AP_L (AP on objects of large scales).

5.1 Ablation Studies

Auxiliary Losses. Since the auxiliary losses can build extra back-propagation paths between the objective functions and the backbone levels, we propose to introduce auxiliary losses to address the dropped AP_L problem. Table 1 shows the ablation experiments. After introducing the auxiliary losses, it is observed that AP_L increases from 50.4% to 51.2% and from 51.6% to 52.2% in the one-stage and two-stage detector, respectively. The improvements indicate that the auxiliary losses can help f_{s_i} , $1 \leq i \leq 3$ own the ability to see all objects and thereby C_5 can carry more effective semantic information of the large-scale objects. Furthermore, we can observe that AP_S drops a lot if simply add the auxiliary losses

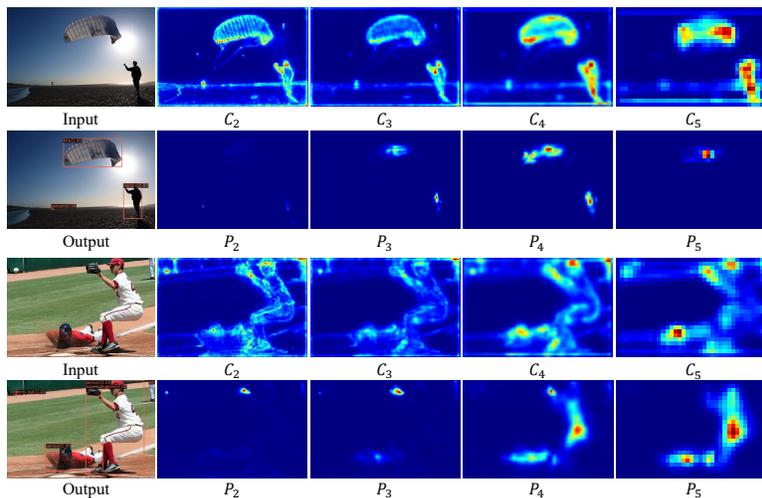


Fig. 5. Visualization of the learned features. The adopted model is Faster R-CNN with ResNet-101. The pictures are selected from MS COCO *val-2017*.

and the original losses linearly. The drops indicate that to some extent, the auxiliary signals will overwrite the original loss signals especially for the small objects whose effective information is the least. To this end, we introduce the uncertainty to the auxiliary losses to scale the auxiliary signals adaptively. It is observed that AP_S improves from 21.2% to 23.3% and from 21.8% to 24.2% in the one-stage and two-stage pipeline, respectively. As a result, the overall detection performance improves with the consistent changes in $\{AP_S, AP_M, AP_L\}$. These improvements well demonstrate the correctness of our speculation and the effectiveness of our method. Furthermore, since the auxiliary predictions do not participate in the model inference phase, the FPS of the detectors will not drop after the introduction of the auxiliary losses.

Feature Grouping. From the perspective of optimization, we have identified the unreasonable operation in the process of building traditional FPN. Specifically, the top-down architecture will make the shallow layers of the backbone network fail to see the large-scale objects. To mitigate the adverse effects caused, we propose to leverage a feature grouping module to construct each feature pyramid level by selecting the feature maps from all backbone stages. Table 2 demonstrates the empirical study. We can observe that the FPN with feature grouping outperforms the traditional FPN by 1.7% and 1.1% in one-stage and two-stage detection framework, respectively. And AP increases with the consistent rises in $\{AP_S, AP_M, AP_L\}$. The result indicates that the feature grouping module has the ability to make each backbone level see all objects through amending the back-propagation paths between objective functions and the backbone network.

Cascade Structure. To achieve more robust improvements by enhancing the space compactness of the homogeneous feature maps, we propose to introduce a cascade feature grouping structure. The experimental results in Table 2 demonstrate the effectiveness of this structure. It is observed that the detection per-

Table 3. The improvements on AP after integrating the cascade feature grouping module into various detection frameworks. The $1\times$, $3\times$ training schedules follow the settings explained in MMDetection [4]. FPN-CFG denotes for applying the cascade feature grouping module into FPN.

Method	Backbone	Schedule	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
<i>one-stage</i>								
RetinaNet [22]	ResNet-101-FPN	$1\times$	38.5	57.6	41.0	21.7	42.8	50.4
RetinaNet [22]	ResNet-101-FPN-CFG	$1\times$	41.2 (+2.7)	60.8	43.8	24.1	45.2	55.2
FreeAnchor [46]	ResNet-101-FPN	$1\times$	40.3	59.0	43.1	21.8	44.0	54.2
FreeAnchor [46]	ResNet-101-FPN-CFG	$1\times$	43.2 (+2.9)	62.0	46.3	24.4	47.4	57.6
ATSS [45]	ResNet-101-FPN	$1\times$	41.5	59.9	45.2	24.2	45.9	53.3
ATSS [45]	ResNet-101-FPN-CFG	$1\times$	43.8 (+2.3)	62.1	47.3	26.8	48.0	57.2
<i>two-stage</i>								
Faster R-CNN [31]	ResNet-101-FPN	$1\times$	39.4	60.1	43.1	22.4	43.7	51.1
Faster R-CNN [31]	ResNet-101-FPN-CFG	$1\times$	42.2 (+2.8)	63.0	45.8	25.5	46.1	55.8
Mask R-CNN [9]	ResNet-101-FPN	$1\times$	40.0	60.5	44.0	22.6	44.0	52.6
Mask R-CNN [9]	ResNet-101-FPN-CFG	$1\times$	43.3 (+3.3)	63.7	47.6	25.7	47.1	56.6
Cascade R-CNN [2]	ResNet-101-FPN	$1\times$	42.0	60.4	45.7	23.4	45.8	55.7
Cascade R-CNN [2]	ResNet-101-FPN-CFG	$1\times$	44.5 (+2.5)	63.1	48.4	26.1	48.5	57.8
Cascade Mask R-CNN [2]	ResNet-101-FPN	$1\times$	42.9	61.0	46.6	24.4	46.5	57.0
Cascade Mask R-CNN [2]	ResNet-101-FPN-CFG	$1\times$	45.4 (+2.5)	63.8	49.4	27.5	49.3	59.5
<i>anchor-free</i>								
FCOS [39]	ResNet-50-FPN	$1\times$	36.6	56.0	38.8	21.0	40.6	47.0
FCOS [39]	ResNet-50-FPN-CFG	$1\times$	39.6 (+3.0)	58.8	42.3	22.9	43.4	51.9
Sparse R-CNN [37]	ResNet-50-FPN	$1\times$	37.9	56.0	40.5	20.7	40.0	53.5
Sparse R-CNN [37]	ResNet-50-FPN-CFG	$1\times$	40.1 (+2.2)	58.7	42.6	22.2	42.6	55.6
FSAF [49]	ResNet-101-FPN	$1\times$	39.3	58.6	42.1	22.1	43.4	51.2
FSAF [49]	ResNet-101-FPN-CFG	$1\times$	42.2 (+2.9)	62.0	44.8	24.3	45.9	56.2
<i>transformer</i>								
Mask R-CNN [27]	Swin-T-FPN	$1\times$	42.7	65.2	46.8	26.5	45.9	56.6
Mask R-CNN [27]	Swin-T-FPN-CFG	$1\times$	46.0 (+3.3)	67.0	50.5	28.8	49.7	59.1
<i>strong baseline</i>								
Cascade Mask R-CNN [2]	ResNeXt-101-64x4d-FPN	$3\times$	46.6	65.1	50.6	29.3	50.5	60.1
Cascade Mask R-CNN [2]	ResNeXt-101-64x4d-FPN-CFG	$3\times$	50.1 (+3.5)	68.6	54.5	32.7	53.7	64.3

formance improves steadily as the number of cascade times increases in both one-stage and two-stage frameworks. Moreover, the improvements of AP always benefit from the consistent improvements of $\{AP_S, AP_M, AP_L\}$.

Visualization of Learned Features. In Figure 5, we visualize the feature maps outputted by the backbone levels (*i.e.*, \mathcal{C}) and pyramid levels (*i.e.*, \mathcal{P}). It is observed that \mathcal{C} contain the semantic information of the whole image, while \mathcal{P} only carries the effective semantics used to detect the objects within the corresponding scale range, indicating that the feature grouping module can well promise the space compactness of the homogeneous feature maps.

5.2 Performance with Various Detection Frameworks.

To further prove the soundness of our principle and the robustness of our approach, we integrate the cascade feature grouping (CFG) structure into various detection frameworks. Table 3 demonstrates the experimental results. For one-stage detectors, our approach consistently improves the baseline frameworks by at least 2.3% AP. For two-stage detectors with pre-defined anchors and ResNet backbone, the baseline frameworks are increased by more than 2.5% AP. Recent academic attention has been geared toward anchor-free detectors and transformer-based backbone networks. We have also made attempts to in-

Table 4. Experimental results on instance segmentation task. The models are trained on the MS COCO *train-2017* split and evaluated on the MS COCO *val-2017* set.

Method	Backbone	Schedule	AP ^{seg}	AP ₅₀ ^{seg}	AP ₇₅ ^{seg}	AP _S ^{seg}	AP _M ^{seg}	AP _L ^{seg}
Mask R-CNN	Swin-T-FPN	1×	39.30	62.20	42.20	20.50	41.80	57.80
Mask R-CNN	Swin-T-FPN-CFG	1×	41.40 (+2.1)	64.50	44.60	21.90	44.60	58.80
Mask R-CNN	ResNet-101-FPN	1×	36.10	57.50	38.60	18.80	39.70	49.50
Mask R-CNN	ResNet-101-FPN-CFG	1×	38.70 (+2.6)	60.80	41.50	19.00	42.00	56.00
Cascade Mask R-CNN	ResNet-101-FPN	1×	37.30	58.20	40.10	19.70	40.60	51.50
Cascade Mask R-CNN	ResNet-101-FPN-CFG	1×	39.40 (+2.1)	61.30	42.60	19.70	42.60	57.10

tegrate the proposed structure into these frameworks. It is observed that the cascade feature grouping structure brings more than 2.2% AP improvements to the anchor-free detectors and the transformer-based detectors. Moreover, we have also trained a strong baseline with multi-scale training, 3× schedule and ResNeXt-101-64x4d backbone. After integrating the cascade feature grouping module into the traditional FPN, the strong baseline is still improved by 3.5% AP. It is worth mentioning that the performance gains are all achieved by consistently boosting the AP of the objects within different scale ranges. The results above together show the necessity and effectiveness that each level of the backbone network should own the ability to look at all objects.

5.3 Instance Segmentation

To verify the generalization ability of our approach, we also apply the cascade feature grouping module on a more challenging instance segmentation task, which requires the prediction of object instances and their per-pixel segmentation mask simultaneously. As shown in Table 4, our method improves AP^{seg} of different detectors from 39.30% to 41.40%, 36.10% to 38.70%, and 37.30% to 39.40%, respectively. Moreover, all the improvements are built upon the consistent increases in $\{AP_S^{seg}, AP_M^{seg}, AP_L^{seg}\}$.

6 Conclusions

This work first identifies the nature of the success of FPN from the perspective of optimization. Based on the principle, we succeed in illustrating the reason why the introduction of FPN will suppress the detection performance of large objects. We further conclude that the key to address the inconsistent changes problem in $\{AP_S, AP_M, AP_L\}$ is to enable each backbone level to look at all objects. Therefrom, we propose to design two strategies to achieve this goal. One is to introduce the auxiliary losses so that the auxiliary signals containing the information of all objects can directly pass through the shallow layers of the backbone network. The other is to integrate the cascade feature grouping structure into the existing FPN, which can also amend the back-propagation paths between the objective functions and the shallow layers of the backbone network. Extensive experiments show the soundness of our principle and the effectiveness of our strategies. Without bells and whistles, our method brings consistent performance improvements to 12 different detection frameworks.

References

1. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms—improving object detection with one line of code. In: Proceedings of the IEEE international conference on computer vision. pp. 5561–5569 (2017)
2. Cai, Z., Vasconcelos, N.: Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019)
3. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al.: Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4974–4983 (2019)
4. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
5. Chen, Q., Wang, Y., Yang, T., Zhang, X., Cheng, J., Sun, J.: You only look one-level feature. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13039–13048 (2021)
6. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
7. Ghiasi, G., Lin, T.Y., Le, Q.V.: Nas-fpn: Learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7036–7045 (2019)
8. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. He, Y., Zhang, X., Savvides, M., Kitani, K.: Softer-nms: Rethinking bounding box regression for accurate object detection. arXiv preprint arXiv:1809.08545 **2**(3) (2018)
12. He, Y., Zhu, C., Wang, J., Savvides, M., Zhang, X.: Bounding box regression with uncertainty for accurate object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2888–2897 (2019)
13. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3588–3597 (2018)
14. Jin, Z., Liu, B., Chu, Q., Yu, N.: Safnet: A semi-anchor-free network with enhanced feature pyramid for object detection. *IEEE Transactions on Image Processing* **29**, 9445–9457 (2020)
15. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems* **30** (2017)
16. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7482–7491 (2018)
17. Kong, T., Sun, F., Tan, C., Liu, H., Huang, W.: Deep feature pyramid reconfiguration for object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 169–185 (2018)

18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25**, 1097–1105 (2012)
19. Li, S., Yang, L., Huang, J., Hua, X.S., Zhang, L.: Dynamic anchor feature selection for single-shot object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 6609–6618 (2019)
20. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *arXiv preprint arXiv:2006.04388* (2020)
21. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2117–2125 (2017)
22. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2980–2988 (2017)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *European conference on computer vision*. pp. 740–755. Springer (2014)
24. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 8759–8768 (2018)
25. Liu, S., Huang, D., Wang, Y.: Adaptive nms: Refining pedestrian detection in a crowd. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6459–6468 (2019)
26. Liu, Y., Wang, R., Shan, S., Chen, X.: Structure inference net: Object detection using scene-level context and instance-level relationships. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 6985–6994 (2018)
27. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030* (2021)
28. Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al.: Mixed precision training. *arXiv preprint arXiv:1710.03740* (2017)
29. Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., Lin, D.: Libra r-cnn: Towards balanced learning for object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 821–830 (2019)
30. Qian, Q., Chen, L., Li, H., Jin, R.: Dr loss: Improving object detection by distributional ranking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 12164–12172 (2020)
31. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **28**, 91–99 (2015)
32. Shrivastava, A., Gupta, A.: Contextual priming and feedback for faster r-cnn. In: *European conference on computer vision*. pp. 330–348. Springer (2016)
33. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 761–769 (2016)
34. Shrivastava, A., Sukthankar, R., Malik, J., Gupta, A.: Beyond skip connections: Top-down modulation for object detection. *arXiv preprint arXiv:1612.06851* (2016)

35. Singh, B., Davis, L.S.: An analysis of scale invariance in object detection snip. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3578–3587 (2018)
36. Singh, B., Najibi, M., Davis, L.S.: Sniper: Efficient multi-scale training. arXiv preprint arXiv:1805.09300 (2018)
37. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al.: Sparse r-cnn: End-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14454–14463 (2021)
38. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10781–10790 (2020)
39. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)
40. Wang, J., Song, L., Li, Z., Sun, H., Sun, J., Zheng, N.: End-to-end object detection with fully convolutional network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15849–15858 (2021)
41. Wu, Y., Chen, Y., Yuan, L., Liu, Z., Wang, L., Li, H., Fu, Y.: Rethinking classification and localization for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10186–10195 (2020)
42. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. <https://github.com/facebookresearch/detectron2> (2019)
43. Yang, W., Zhang, T., Yu, X., Qi, T., Zhang, Y., Wu, F.: Uncertainty guided collaborative training for weakly supervised temporal action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 53–63 (2021)
44. Zhang, D., Zhang, H., Tang, J., Wang, M., Hua, X., Sun, Q.: Feature pyramid transformer. In: European Conference on Computer Vision. pp. 323–339. Springer (2020)
45. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9759–9768 (2020)
46. Zhang, X., Wan, F., Liu, C., Ji, X., Ye, Q.: Learning to match anchors for visual object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021)
47. Zhao, G., Ge, W., Yu, Y.: Graphfpn: Graph feature pyramid network for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2763–2772 (2021)
48. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
49. Zhu, C., He, Y., Savvides, M.: Feature selective anchor-free module for single-shot object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 840–849 (2019)
50. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9308–9316 (2019)