FEW-SHOT OBJECT DETECTION WITH LOCAL CORRESPONDENCE RPN AND ATTENTIVE HEAD

Jian Han *Yali Li** *Shengjin Wang*

Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

ABSTRACT

Existing object detection methods rely heavily on a large number of annotated bounding boxes, which is expensive to collect. In this paper, we propose a novel few-shot object detection method named GCN-FSOD. Intending to find informal local correspondence to fully explore cues of novel classes, we propose the local correspondence region proposal network (lcRPN) and the attentive detection head for few-shot detection. Taking features from the support-query image pair as inputs, lcRPN generates region proposals by mining fine-grained local correspondence with the help of GCNs. Then the proposed attentive head performs precise detection. We conduct extensive experiments on the wildly adopted MS-COCO benchmark. The proposed GCN-FSOD brings significant performance gains and outperforms the state-of-the-art by a large margin (1.7% mAP for 10-shot).

***Index Terms*—** few-shot detection, GCNs, attention

1. INTRODUCTION

Object detection aims at localizing and classifying objects from given images, which is a fundamental problem for many higher-level computer vision tasks. With the rise of deep convolution networks in recent years, some representative works, *e.g.*, SSD[1], YOLO[2], Faster R-CNN[3], and Mask R-CNN[4] have achieved huge success in objection detection. However, these methods rely heavily on a large amount of annotated bounding boxes. These large-scale annotated data are very expensive to collect. Besides, there are some scenarios where the examples are scarce, *e.g.*, endangered animals or some medical data. Therefore, few-shot object detection (FSOD) which models novel categories from a few annotations becomes very important.

Provided with a few annotated bounding boxes, few-shot object detection tries to detect objects from novel classes with the aid of base categories that contain sufficient labeled

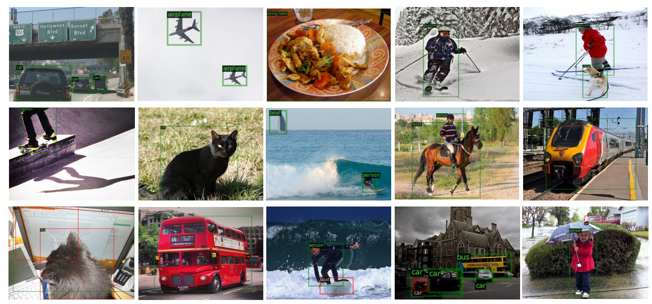

Fig. 1: *Visualization of detection results for GCN-FSOD.*

data. It is more challenging than few-shot image classification since we have to localize the objects in a cluttered background apart from recognizing the objects of novel categories. Some methods [5, 6, 7] attempt to solve this issue by attaching a meta learner to generic object detection pipelines. The meta learner offers category attentive vectors to guide the learner (detector) to detect novel class objects. But these methods are still stuck with visual cues sparsity when generating category attentive vectors for novel classes. Besides, they assume that the Region Proposal Network (RPN) trained in a class-agnostic manner could find potential region proposals for novel categories. This assumption is unreliable because RPN has very few novel class examples for reference and over-fits easily.

In this work, we propose a novel category-aware RPN named lcRPN to generate region proposals for specific categories. lcRPN detect objects in the query image from the same category with the support object patches. To fully explore the limited visual cues of novel classes, we introduce a learnable local and global correspondence mechanism in lcRPN to recall latent objects. With the help of Graph Convolution Networks (GCNs), local features from the query and support are matched and aggregated automatically. More fine-grained features are exploited to improve the proposal generation. Considering the aggregation operations of GCNs may smooth out the sharp edges information, which is crucial to the bounding box regression. We decouple the classification and regression in lcRPN to learn task-specific patterns. In the second detection stage, we propose a novel attentive detection head to further decouple the classification and box regression. More fine-grained local cues are exploited

This work was supported by the state key development program in 14th Five-Year under Grant No.2021YFF0602103, No.2021YFF0602102, and No.2021QY1702. We also thank for the research fund under Grant No.2019GQG0001 from the Institute for Guo Qiang, Tsinghua University.
*The corresponding author is Yali Li, E-mail: liyali13@tsinghua.edu.cn

through the attentive manner to supplement the visual sparsity. The main contributions are summarized as follows:

- We propose a novel local correspondence RPN (lcRPN) to generate region proposals. More fine-grained visual information is exploited to search for latent objects. The decoupling design learns task-aware patterns.
- We propose an attentive detection head that mines local and global consistency for classification while exploiting class-attentive features for regression.
- Extensive experiments are conducted on the MS-COCO benchmark. The proposed GCN-FSOD reaches the highest mAP and exceeds the state-of-the-art methods by large margins (1.7% mAP for 10-shot).

2. RELATED WORKS

General Object detection. Basically, deep learning based detection can be divided into three groups: two-stage ones such as [8, 9, 3, 4], one-stage ones such as [1, 2, 10, 11, 12] and anchor-free methods such as [13, 14, 15, 16]. Two-stage methods, as the name suggests, detect instances through two progressive stages where in the first stage a region proposal network is exploited to generate potential object candidate regions and get corresponding feature representations. In the second stage, they carry out more precise classification and bounding box regression. In contrast to two-stage methods, one-stage detectors directly predict labels and regress bounding boxes on the feature map.

Few-shot Object detection. LSTD[17] proposes a combining model of SSD[1] and Faster R-CNN[3]. The multi-scale features in SSD compensate for the lack of scale diversity caused by limited data. The background depression and transfer knowledge regularization alleviates over-fitting. TFA[18] divides Faster R-CNN into a feature extractor and a bounding box predictor. Only the bounding box predictor is fine-tuned. FSRW[5] and Meta R-CNN[6] are based on YOLO and Faster R-CNN respectively. A meta learner is build to extract class attentive vectors. They employ the vectors to reweight the meta-features and perform specific category detection. Attention RPN presented in [19] uses support information to filter out most background boxes and mismatching categories' objects. But the global correlation in Attention RPN may suffer spatial scale misalignment and the local correspondence between objects is neglected.

3. METHODS

In few-shot object detection, the training data is divided into two groups: $D_{base} = \{(x_i^{base}, y_i^{base})\}_{i=1}^{n_{base}} \sim P_{base}$ contains sufficient samples in each base class; $D_{novel} = \{(x_i^{novel}, y_i^{novel})\}_{i=1}^{n_{novel}} \sim P_{novel}$ contains very few samples in each novel class. The FSOD detector $h(;\theta)$ is firstly trained on D_{base} and then fine-tuned on D_{novel}. During the inference stage, $h(;\theta)$ aims to localize and classify objects drawn from P_{novel}. The inputs of the proposed GCN-FSOD are composed of positive and negative support-query image pairs *i.e.*, (Q_i, S_j^+) or (Q_i, S_j^-). In particular, (Q_i, S_j^+) consists of one query image and K object patches from the same category while (Q_i, S_j^-) consists of one query image and K object patches from other categories.

3.1. Local Correspondence RPN

As shown in Fig. 2, Query-support image pair (Q_i, S_j) sampled from D_{base} and D_{novel} are fed to lcRPN to search for the correspondence. In this case, objects consistent with S_j in the query image are selected while others are suppressed. In particular, we compute the local correspondence between the feature map of the query and that of the support through a GCN network. The local correspondence then is utilized to generate proposal objectnesss. We denote the feature map of S_j as $\boldsymbol{S}_f \in \boldsymbol{R}^{h\times w\times C}$ and the feature map of Q_i as $\boldsymbol{Q}_f \in \boldsymbol{R}^{H\times W\times C}$. The features of S_j and Q_i are spatially split into multiple vectors as $\boldsymbol{V}^S = \{\boldsymbol{v}_1^S, \boldsymbol{v}_2^S, ..., \boldsymbol{v}_{h\times w}^S\}$ and $\boldsymbol{V}^Q = \{\boldsymbol{v}_1^Q, \boldsymbol{v}_2^Q, ..., \boldsymbol{v}_{H\times W}^Q\}$, where $\boldsymbol{v} \in \boldsymbol{R}^C$. We build a graph with the vectors of $\boldsymbol{V} = \boldsymbol{V}^S \cup \boldsymbol{V}^Q$ as nodes. The adjacency matrix $\boldsymbol{A}$ represents visual correlations among the local feature vectors in the graph. It is calculated by the euclidean distance with Gaussian kernel as in Eq. 1. T is the temperature parameter of the kernel.

$$\boldsymbol{A} = \{a_{ij} = \exp(-\frac{\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2}{\|\boldsymbol{v}_i\|\|\boldsymbol{v}_j\|}/T) | \boldsymbol{v}_i, \boldsymbol{v}_j \in \boldsymbol{V}\} \tag{1}$$

$$\boldsymbol{H}^{l+1} = \sigma(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{H}^l\Theta^l), l = 0, 1 \tag{2}$$

The GCNs in lcRPN consists of two graph convolution layers. The propagation rule to perform convolution on the graph following Eq. 2, which H^l represents the activation in the l^{th} layer and $\Theta \in \boldsymbol{R}^{C^{l-1}\times C^l}$ denotes a trainable weight matrix for layer l with C^l corresponding to the number of learned filters. For the first layer, $\boldsymbol{H}^{l=0} = \boldsymbol{V}$. $\sigma(\cdot)$ denotes a nonlinear activation function and here is ReLU. $\boldsymbol{D}$ is the degree matrix of $\boldsymbol{A}$ which normalizes $\boldsymbol{A}$ to ensure the scale stability of the features. The constructed graph aggregates node features with neighborhood nodes, *i.e.* visual similar local features. As a result, we obtain the local correspondence $\boldsymbol{L} = \boldsymbol{H}^{l=2}$ on the query feature $\boldsymbol{Q}_f$ activated by itself and the local support feature. The reshape operation is applied on the local correspondence $\boldsymbol{L}$ to restore the shape of $\boldsymbol{R}^{H\times W\times C}$. After that, it becomes $\boldsymbol{Q}_{cls}$ for classification and then produces objectness of proposals.

The local correspondence $\boldsymbol{L}$ are obtained by aggregating neighborhood node features, where the sharp edges information is smoothed out. To preserve the critical sharp edges information, we decouple the classification and regression into

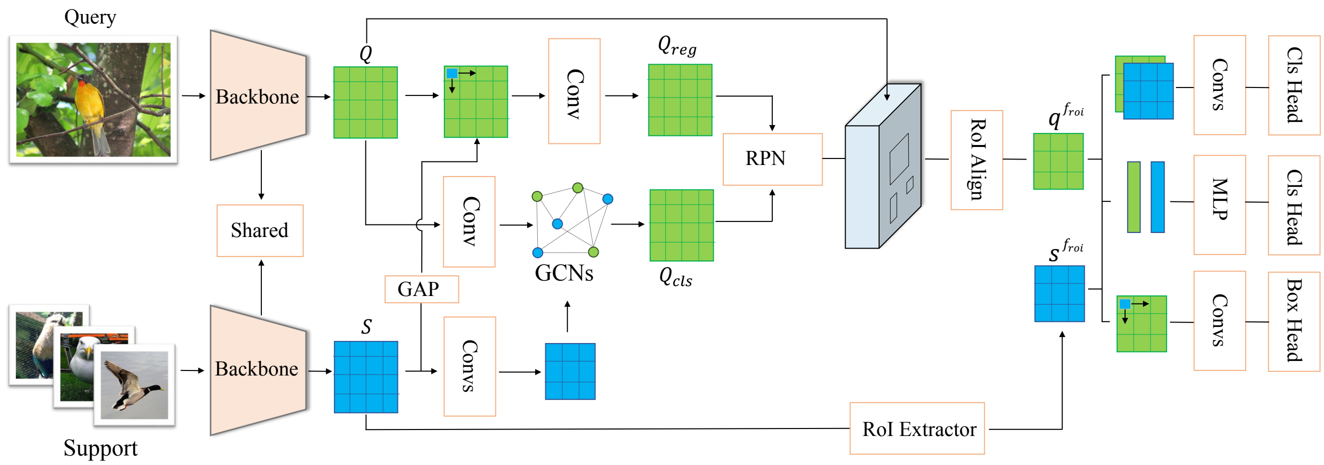

Fig. 2: *The framework of GCN-FSOD. It takes query-support image pairs as inputs. lcRPN generates region proposals by computing local correspondence and learn task-specific patterns with decoupling. Then the proposed attentive head perform precise detection in the second stage.*

two isolated branches to learn task-specific patterns. For the bounding boxes regression, the feature map $\boldsymbol{S}_f$ of support images are globally average pooled as the class attentive vector $\boldsymbol{s}_f$. Then the feature map $\boldsymbol{Q}_f$ of the query are activated by channel-wise multiplication with $\boldsymbol{s}_f$. Through the class activation operation, the meta feature map $\boldsymbol{Q}_f$ becomes class-aware. After that, it experiences a 3×3 convolution layer to become $\boldsymbol{Q}_{reg}$ for regression and then produces bounding box offsets for the predefined anchors.

3.2. Attentive Detection Head

With the RoI Aligh and RoI extractor, we obtain RoI features of $\{\boldsymbol{s}^{f_{roi}}, \boldsymbol{q}_1^{f_{roi}}, ..., \boldsymbol{q}_M^{f_{roi}}\}$. The detection head in Faster R-CNN[3] whose input is the single RoI feature instead of the RoI feature pair is unusable. We propose an attentive detection head consisting of the dual-level classification and the category-attentive regression as illustrated in Fig.2. The dual-level classification incorporates the global matching and local matching. The global matching is implemented by globally average pooling RoI features $(\boldsymbol{s}^{f_{roi}}, \boldsymbol{q}_i^{f_{roi}})$, concatenating results as global vector $\boldsymbol{q}_i^g$. As for the local matching, we calculate the local relevance by point-wise multiplication $\boldsymbol{s}^{f_{roi}} \odot \boldsymbol{q}_i^{f_{roi}}$. Then we perform spatial summation and flatten the results into the local vector $\boldsymbol{q}_i^l$. After that, we embed $\boldsymbol{q}_i^g$ and $\boldsymbol{q}_i^l$ into another space and perform classifications. The binary classification distinguishes RoIs of $(\boldsymbol{s}^{f_{roi}}, \boldsymbol{q}_i^{f_{roi}})$ come from the same category or not. The global matching considers the consistency between the global features while the local matching considers the local correlations between RoI features. The local matching fully explores the limited visual cues for novel categories. The category-attentive box regression aims to localize the object precisely. In particular, the RoI feature $\boldsymbol{s}^{f_{roi}}$ is globally average pooled as the class attentive vector. Then the class attentive vector serves as the kernel to slide on the query's RoI feature $\boldsymbol{q}_i^{f_{roi}}$ to emphasize the class-related information of a specific category. After that, the attentive RoI feature experiences several convolutions and mean pooling layers to regress bounding boxes.

4. EXPERIMENTS

4.1. Datasets and Settings

Datasets and setups. We evaluate our method on the MS-COCO[21] benchmark. MS-COCO consists of 80 object categories including the 20 categories in PASCAL VOC[22]. We follow the same few-shot setting in [5, 6]. In particular, 60 classes disjointed with PASCAL VOC are used as the base classes and the rest 20 classes which are the same with PASCAL VOC are taken as novel classes. We report the "10-shot 20-way" and "30-shot 20-way" results on MS-COCO. The mean Average Precision of novel categories is presented.

Implementation Details. We use ResNet-50 pretrained on ImageNet[23] as the backbone and use the SGD optimizer with a mini-batch size of 8 on four TITAN XP GPUs. We train 120k iterations with the learning rate of 0.002 and divide it by 10 in the last 8000 iterations during base training. For novel class fine-tuning, we train 2000 iterations with the learning rate of 0.001 and divide it by 10 in the last 700 iterations for 10-shot while the fine-tuning iterations are 3000 and the learning rate drops in the last 1000 iterations for 30-shot. Support image patches are cropped from annotated images and resized to 320 with zero paddings.

4.2. Comparison with the State-of-the-Art Methods

The MS-COCO benchmark is very challenging due to image scale variations, small objects, and cluttered backgrounds. The few-shot evaluation results on MS-COCO are illustrated in Tab. 1. Our method reaches 12.8% mAP, 24.0% AP50, and 12.3% AP75 for 10-shot, exceeding meta learning methods FSRW[5], Meta R-CNN[6], SRR[7] (11.3% mAP) and fine-tuning methods LSTD[17], TFA[18], Attention RPN[19], Re-

Table 1: *Few-shot object detection evaluation on MS-COCO. We report the mean Averaged Precision and mean Averaged Recall on the 20 novel classes of COCO.*

		Backbone	Average Precision						Average Recall					
Shots	Method		0.5:0.95	0.5	0.75	S	M	L	1	10	100	S	M	L
10	LSTD [17]	VGG16	3.2	8.1	2.1	0.9	2.0	6.5	7.8	10.4	10.4	1.1	5.6	19.6
	FSRW [5]	DarkNet	5.6	12.3	4.6	0.9	3.5	10.5	10.1	14.3	14.4	1.5	8.4	28.2
	MetaDet* [20]	VGG16	7.1	14.6	6.1	1.0	4.1	12.2	11.9	15.1	15.5	1.7	9.7	30.1
	Meta R-CNN* [6]	ResNet101	8.7	19.1	6.6	2.3	7.7	14.0	12.6	17.8	17.9	7.8	15.6	27.2
	TFA* w/fc [18]	ResNet101+FPN	9.1	17.3	8.5	–	–	–	–	–	–	–	–	–
	TFA* w/cos [18]	ResNet101+FPN	9.1	17.1	8.8	–	–	–	–	–	–	–	–	–
	Attention RPN [19]	ResNet50	11.1	20.4	10.6	–	–	–	–	–	–	–	–	–
	GCN-FSOD (Ours)	ResNet50	**12.8**	**24.0**	**12.3**	**2.7**	**13.8**	**22.4**	**19.6**	**27.6**	**27.6**	3.2	**25.2**	**48.0**
30	LSTD [17]	VGG16	6.7	15.8	5.1	0.4	2.9	12.3	10.9	14.3	14.3	0.9	7.1	27.0
	FSRW [5]	DarkNet	9.1	19.0	7.6	0.8	4.9	16.8	13.2	17.7	17.8	1.5	10.4	33.5
	MetaDet* [20]	VGG16	11.3	21.7	8.1	1.1	6.2	17.3	14.5	18.9	19.2	1.8	11.1	34.4
	Meta R-CNN* [6]	ResNet101	12.4	25.3	10.8	2.8	11.6	19.0	15.0	21.4	21.7	**8.6**	20.0	32.1
	TFA* w/fc [18]	ResNet101+FPN	12.0	22.2	11.8	–	–	–	–	–	–	–	–	–
	TFA* w/cos [18]	ResNet101+FPN	12.1	22.0	12.0	–	–	–	–	–	–	–	–	–
	GCN-FSOD (Ours)	ResNet50	**15.2**	**27.7**	**14.7**	**3.7**	**16.0**	**25.6**	**21.2**	**30.4**	**30.4**	5.6	**28.4**	**51.6**

Table 2: *Ablation studies on local correspondence RPN.*

Method	mAP	AP50	AP75	R1	R10
global feature kernel	11.9	21.6	11.9	18.8	24.8
lcRPN (w/o decouple)	12.4	22.7	12.2	19.2	26.4
lcRPN (Ours)	12.8	24.0	12.3	19.6	27.6

Table 3: *Ablation studies on the attentive head.*

C_g	C_l	R_s	R_c	mAP	AP50	AP75	R1	R10
✓		✓		10.6	20.3	9.9	18.4	25.6
	✓	✓		10.4	20.5	9.5	18.0	25.6
✓	✓	✓		12.2	22.6	12.2	19.2	26.8
✓	✓		✓	12.8	24.0	12.3	19.6	27.6

tentive RCNN[24] (10.5% mAP) by a large margin. Note that our method achieves better performance even with a weaker backbone compared to Meta R-CNN[6] and TFA[18]. Similar performance advantages could be observed for the 30-shot setting, the proposed GCN-FSOD achieves 15.2% mAP, 27.7% AP50, and 14.7% AP75, outperforming TFA[18] by 3.1%, 5.7%, and 2.7% respectively. These considerable improvements indicate that our proposed method is very effective for few-shot object detection. We attribute the large performance gains of low shot to the local correspondence RPN and the attentive detection head. Visual cues sparsity could be alleviated by fully explored the local visual features' correspondence. In addition to Average Precision, the proposed GCN-FSOD obtains the highest Average Recall when selecting different numbers of predictions as shown in Tab. 1. It suggests that GCN-FSOD could recall more latent objects.

4.3. Ablation Studies

We perform a series of ablation studies on the MS-COCO benchmark with 10 shots to verify the partial effectiveness of the proposed method.

Local correspondence RPN. We calculate the correspondence and aggregate similar local features by learnable GCNs. We compare with the globally pooled support feature to slide on the query feature, as proposed in Attention RPN[19]. The results are shown in Tab. 2, the adopted lcRPN which calculates local correspondence by GCNs could yield 0.5% mAP and 1.1% AP50 improvements. It suggests exploiting local correspondence to generate proposals is effective. Another design in lcRPN is the decoupling of classification and box regression. Compared to lcRPN without task disentanglement, the decoupling brings extra 0.4% mAP and 1.3% AP50 improvements. It demonstrates that the decoupling designing is very helpful to learn task-specific patterns.

Attentive detection head. The attentive detection head consists of the global matching, the local matching, and the category-attentive box regression. Except for the mentioned subheads, we also implement the regression head predicting on the stack RoI features of support and query (denote as R_s). The ablation results are listed in Tab.3. It can be seen that the global matching head is better than the local matching head, while the combination of them generates the best results. It demonstrates they match different level features and could complement each other. Compared to R_s, the category-attentive regression head R_c generates 0.6% mAP and 1.4% AP50 gains. It demonstrates the explicit class attention is better than implicit feature stacking.

4.4. Visualization and Analysis.

We visualize the qualitative detection results on the MS-COCO dataset in Fig. 1. Here success cases are marked in green boxes while failure cases in red boxes when detecting novel category objects. It can be seen that the proposed method could successfully detect novel objects with relatively precise bounding boxes. The failure cases mainly are misclassifying novel objects as similar objects.

5. CONCLUSION

Few-shot object detection alleviates the difficulty of collecting large-scale expensive annotated data. In this paper, we propose a novel few-shot object detection approach named GCN-FSOD. It consists of lcRPN and attentive detection head intending to find informal local correspondence to supplement the visual sparsity. Extensive experiments demonstrate the effectiveness of our method. GCN-FSOD outperforms the state-of-the-art methods with a large margin and serves as a strong pipeline for few-shot object detection.

6. REFERENCES

[1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European Conference on Computer Vision (ECCV)*, 2016.

[2] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[5] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, "Few-shot object detection via feature reweighting," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[6] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta R-CNN : Towards general solver for instance-level low-shot learning," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[7] C. Zhu, F. Chen, U. Ahmed, Z. Shen, and M. Savvides, "Semantic relation reasoning for shot-stable few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8782–8791.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[9] R. Girshick, "Fast R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[11] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4203–4212.

[12] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10781–10790.

[13] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[14] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6569–6578.

[15] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 850–859.

[16] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9657–9666.

[17] C. Hao, W. Yali, W. Guoyou, and Q. Yu, "LSTD: A low-shot transfer detector for object detection," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[18] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," in *International Conference on Machine Learning (ICML)*, July 2020.

[19] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-rpn and multi-relation detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4013–4022.

[20] Y.-X. Wang, D. Ramanan, and M. Hebert, "Meta-learning to detect rare objects," in *IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[21] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014.

[22] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *International Journal of Computer Vision (IJCV)*, 2015.

[23] 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," *International Journal of Computer Vision (IJCV)*, 2015.

[24] Z. Fan, Y. Ma, Z. Li, and J. Sun, "Generalized few-shot object detection without forgetting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4527–4536.