Local-to-global Point Supervised Object Detector via Aggregation of Discriminative Parts *

Yidan Zhang[1,2,3,†], Yingyan Hou[1,2,3,†], Xiaoxuan Liu [1,2,3,‡], Xiaohe Li [1,2,3], Fangli Mou [1,2,3], Peirong Zhang [1,2,3], Xiyu Qi [1,2,3], Jie Jia [1,2,3], Lei Wang [1,2,3], and Xinyu Zhao [1,2,3]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, 100190, China
[2] Key Laboratory of Target Cognition and Application Technology (TCAT), Aerospace Information Research Institute, Beijing, 100190, China
[3] Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Beijing, 100190, China

Abstract. Advanced fully supervised detectors benefit from abundant bounding-box annotations which accurately cover multi-scale objects. However, for point supervised object detectors (PSOD), each object is annotated by a single point without the information of scales. Some scholars have started to represent the scale of the objects through point-to-box regression, yet the accuracy is constrained by manual heuristic algorithms or local part activations. In this paper, we propose a Local-to-global Point Supervised Object Detector called LPSNet, which can adaptively generate globally aware pseudo bounding boxes. Initially, Point-level Prediction (PLP) precisely identifies the object's location at a point level. Subsequently, Box-level Prediction with Aggregation of Discriminative Parts (BPAP) dynamically performs regression from points to component-level proposals, and consolidate those part proposals into global ones. Finally, under the supervision of the pseudo proposals, LPSNet with region proposal network and detection head attached obtains detection results. Extensive experiments on the MSCOCO datasets underlines the superiority of our approach, outperforming other available PSOD methodologies.

Keywords: Object Detection · Point Supervised Object Detection

1 Introduction

Over the past decade, we have witnessed the substantial progress of visual object detection, which is attributed to the representation capacity provided by neural network [21, 34, 17, 42], the effectiveness and efficiency of detection

* Supported by the Key Laboratory Fund of Chinese Academy of Sciences under Grant CXJJ-23S032 and CXJJ-22S032. † Equal Contribution. ‡ Corresponding Author.

EAI MLICOM 2023, December 17, Beijing, People's Republic of China

DOI 10.4108/eai.17-12-2023.2342502

frameworks [2, 46, 48, 16] and the supervision from considerable manual-labeled datasets [7, 32, 11, 25, 45].

Thanks to meticulously crafted manual bounding-box annotations, detectors have been able to capitalize on spatial scales. Drawing inspiration from the advancements in multi-scale anchor methodologies [30, 23, 22], recent innovations in multi-scale feature assignment [39, 20, 52, 49, 50] and geometric relation modeling [18, 8, 6] have been introduced to significantly enhance detection performance.

However, these approaches heavily rely on a wealth of high-quality bounding-box annotations, requiring an approximate time of 34.5 seconds per image for manual labeling [28]. In contrast, using a point-click approach, which provides a point tag (taking only 1.87 seconds per image) containing spatial location and classification label, has emerged as a more efficient alternative to bounding-box annotations. The adoption of point labels for training advanced detectors has gained popularity. Nevertheless, point-level annotation offers only approximate location information, lacking precise object scale data. Consequently, detectors trained solely on this form of supervision exhibit lower accuracy in their object recognition capabilities.

In addressing the aforementioned challenges, a series of studies believe the aim of point supervised object detection is to find the scale of object by implementing a point-to-box regressor. P2B [5] empirically generates multi-scale and -ratios anchors on feature pyramid network [22] to cover the possible extents of objects. However, this kind of method needs lots of fine-tuning on hyperparameter when faced with different datasets [25, 11] with different distribution of object scales, which causes low generalization. In efforts to circumvent the challenges stemming from heuristic methodologies, some studies [51, 19, 15] employ Class Activation Mapping (CAM) to dynamically perceive object scales. They encounter issues related to local activations, leading to the generation of pseudo box-level labels that are constrained in their ability to comprehensively cover object instances.

To generate appropriate, globally aware box-level pseudo labels, we propose a Local-to-global Point Supervised Object Detector called LPSNet. LPSNet constructs a backbone network based on Vison Transformer (ViT) [10] to provide high-order semantic features for subsequent object detection. Subsequently, two modules, Point-level Prediction (PLP) and Box-level Prediction with Aggregation of Discriminative Parts (BPAP), are proposed to convert the features into object potential points and then into box-level proposals. PLP initially confirms the location of the object at the point level. BPAP is then responsible for adaptively regressing from points to component-level proposals and aggregating local parts into global proposals. Finally, under the supervision of the global pseudo-proposals, LPSNet with region proposal network (RPN) and detection head attached can be trained and inferred.

Contributions of this paper are as follows:

- We propose a point supervised detector LPSNet that can adaptively regress local points to globally aware box-level proposals.

- LPSNet is trained in an end-to-end fashion, based on generating the locations of object points in PLP, obtaining box-level proposals by aggregating discriminable features in BPAP.
- Only with single points as supervision, LPSNet achieves advanced precision on commonly used benchmarks.

2 Related Work

2.1 Box-Supervised Object Detection

Box-supervised object detection represents a conventional method for recognizing objects, providing the network with specific category labels and box coordinates. Single-stage detectors, such as YOLO [41], SSD [13], and RetinaNet [23], infer object classifications and bounding-box adjustments by using predetermined anchor settings. On the other hand, two-stage detectors employ various techniques, like RPN in Faster R-CNN [30], to propose boxes and subsequently perform classification and bounding-box adjustments on a subset of these proposed boxes. Transformer-based detectors (DETR [3], Deformable-DETR [53], and Swin-Transformer [26]) leverage global information to enhance overall object representation. Sparse R-CNN [35] integrates the strengths of transformers and CNNs in a sparser detection approach. Nonetheless, creating annotations at the box level incurs substantial costs.

2.2 Image-Supervised Object Detection

To solve the problem of high cost of fully supervised labeling, weakly supervised learning has gradually attracted the attention of scholars. Traditional weakly supervised object detection (WSOD) methods mainly focus on how to train the detector with only image-level annotation. The core idea is how to select or refine more accurate bounding boxes from a bunch of rough object proposals. According to whether object proposal algorithms are adopted as external modules independent of the detectors, WSOD methods can be divided into two-stage fashion [1, 40, 44, 36] and end-to-end fashion [38, 33, 27].

The two-stage fashion as the major approach formulates WSOD as an multiple instance learning (MIL) procedure, which treats each training image as a "bag" and iteratively selects positive instances from each bag when learning detectors. WSDDN [1] builds the first MIL network by integrating an MIL loss into a deep network. Online instance classifier refinement [9, 37, 40, 44] is proposed to select high-quality instances which are treated as pseudo objects to refine the instance classifier. Proposal cluster learning [36] further improves object localization based on proposal clustering, which prevents networks from concentrating on object parts.

Recent methods [38, 33, 27] attempt to break the two-stage WSOD routine with end-to-end approaches. The WeakRPN method [38] utilizes the object contours in convolutional feature maps to generate proposals used to trained an

RPN. In [33], an RPN is trained using the pseudo objects predicted by the weakly supervised detector in a self-training fashion. Wile SPE method [27] further replaced RPN by introducing the attention mechanism in visual transformer and sets the first solid baseline for end-to-end WSOD with sparse proposals.

Relying solely on the category information of the specific object within the image as overall supervision proves to be quite challenging, resulting in a less than optimal object detection accuracy.

2.3 Point-Supervised Object Detection

Unlike WSOD, point supervised object detection (PSOD) methods aim to train the detector with point-level annotation which includes more object localization information. So the PSOD methods are more likely to achieve the performance upper bound from box-supervised object detection. Besides, the annotation cost of PSOD (1.87 seconds per image) is close to that of WSOD (1.5 seconds per image) according to statistics [14, 29] performed on VOC dataset, which is also acceptable.

P2B [5] is known for its empirical generation of multi-scale and multi-ratio anchors within the feature pyramid network [22] to encompass potential object extents. However, this approach necessitates extensive fine-tuning of hyperparameters, particularly when confronted with diverse datasets [25, 11] showcasing varying distributions of object scales, ultimately resulting in reduced generalization. In an attempt to overcome the limitations associated with heuristic methodologies, the Class Activation Mapping (CAM) methods are proposed to dynamically gauge object bounding boxes [51, 19, 15]. However, they face challenges regarding local activations, leading to the generation of pseudo box-level labels that have limitations in adequately encompassing object instances.

To alleviate the above problems, we innovatively propose LPSNet, which can adaptively generate globally aware proposals.

3 Methodology

The Local-to-global Point Supervised Object Detector (LPSNet) predicts pseudo-boxes with point annotations to train a customized ViT structure-based detectors. In LPSNet, Point-level Prediction (PLP) module is first applied to generate the locations of potential object points, and then Box-level Prediction with Aggregation of Discriminative Parts (BPAP) module obtains local box-level proposals, and get globally aware box-level results by aggregating discriminative attention parts. The complete architecture of LPSNet is shown in Fig. 1.

3.1 Point-level Prediction

LPSNet uses modified ViT as the backbone network, including Q shared layers and M unshared ones. Specifically, this backbone is composed of a cascade of

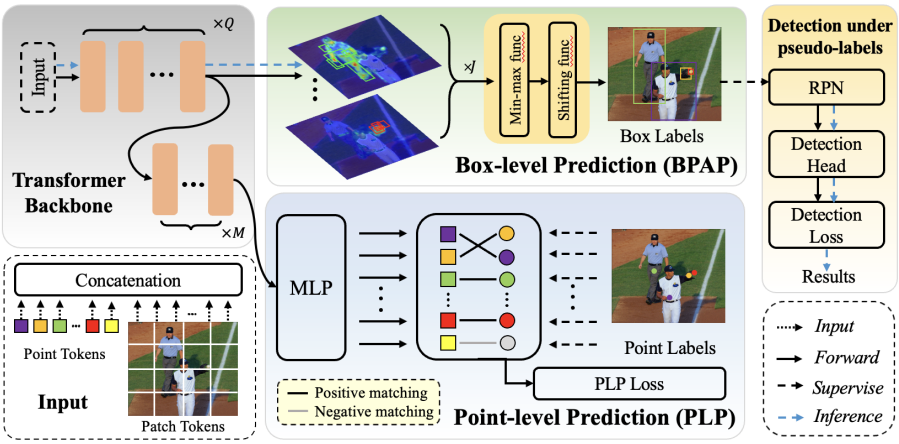

Fig. 1. The architecture of LPSNet. With the support of BLP and BPAP, box-level pseudo-labels are generated and the detector is trained simultaneously. PLP Loss and Detection Loss represent $\mathcal{L}_{PLP}$ and $\mathcal{L}_{DET}$ in Eq.2 and 3 respectively.

self-attention blocks, each of which has a self-attention layer and a multi-layer perception layer.

Backbone Input. The input of the backbone network consists of two parts patches tokens and points tokens concatenated. To obtain inputs that can be processed by the self-attention blocks, N D-dimensional patches tokens $\mathbf{t}^{\mathbf{patch}} = \{t_n^{patch}\}_{n=1}^{N} \in \mathbb{R}^{N\times D}$, are formed from a $W \times H$ image passing through a convolutional layer with $S \times S$ kernel size and stride, where $S = 16$ normally. Additionally, a randomly initialized vector, called a position embedding, is added to each patch token to learn spatial relationships. To represent the potential K object points, we initialize K vectors called points tokens $\mathbf{t}^{point} = \{t_i^{point}\}_{i=1}^{K} \in \mathbb{R}^{K\times D}$ concatenating to patches tokens.

Point Prediction. The points tokens embedded in ViT are trained to predict object points, Fig. 1. In the PLP module, each point token t_{pi} predicts a classification score $s_i \in \mathbb{R}^{1\times C}$ (C is the object category number) by passing three perception layers equipped with ReLU activation and shortcut connection. Using t_{pi} as input, another three MLP with same structure are used to predict a point location $p_i = \{x_i, y_i\} \in \mathbb{R}^{1\times 2}$, which denote normalized coordinates.

Following some transformer-based detectors [12, 3], we utilize the one-to-one matching loss to assign predicted points to the ground-truth points. The predicted points without assignment are categorized as negatives. The point matching loss of a matched ground-truth point is defined as positive loss

$$\mathcal{L}_{pos} = \lambda_1 \mathcal{L}_{cls}(s_i, y_j) + \lambda_2 \mathcal{L}_{L_1}(p_i, g_j) \tag{1}$$

where $\mathcal{L}_{cls}(\cdot)$ denotes the focal loss [24], and $\mathcal{L}_{L_1}(\cdot)$ the point location loss. The y_j and g_j respectively denote the one-hot label of the j-th ground-truth

point and its coordinate in a single image. Additionally, λ_1 and λ_2 are the regularization factors.

The negative points only calculate $\mathcal{L}_{neg} = \lambda_1 \mathcal{L}_{cls}(s_i, o)$, where o is the one-hot background label. Overall, the loss of point prediction is termed as

$$\mathcal{L}_{PLP} = \mathcal{L}_{pos} + \mathcal{L}_{neg} \tag{2}$$

3.2 Box-level Prediction with Aggregation of Discriminative Parts

In the previous section, the J ground-truth points all matched the positive sample points tokens. Next, we sample and aggregate Vit's Q-layer features to generate local box-level proposals.

Attention Maps of Local Activations. For the j-th positive sample point token, by extracting the relevant parts of point and ground truth from the $(N+K)\times(N+K)$ attention matrix inherent in ViT, the attention maps $\mathbf{A}_j$ containing object local activation information can be obtained.

Aggregation of Discriminative Parts. For the attention map A_j, we find discriminating parts $D_j = \mathcal{T}(A_j, \delta_1, \delta_2) = \{d_1, d_2, ..., d_z\}$, where $\mathcal{T}(\cdot)$ denotes a thresholding function [27] with a fixed threshold δ_1 to binarize each attention map and δ_2 to filter the noise. D_j is a collection of local bounding-box proposals containing part areas of the object instance.

We design a transformation function $\mathcal{F} : D_j \rightarrow b_j$ to obtain a globally aware bounding-box proposal from the local ones. The practice of this function is detailed as following:

- **Step1: Min-max function.** Generate the minimum bounding rectangle of all part proposals in the D_j, Fig. 2(a).
- **Step2: Shifting function.** b_j is obtained by moving the center point of the minimum bounding rectangle of the previous step to the j-th matched predicted point, Fig. 2(b).

In the end, globally aware box-level proposals are generated as $\mathbf{B} = \{b_j\}_{j=0}^{J}$.

3.3 Detection under box-level pseudo-labels

Under the supervision of bounding-box labels $\mathbf{B}$, we combine RPN and detection head to achieve object detection tasks.

In the training phase, the overall loss is defined as

$$\mathcal{L} = \mathcal{L}_{PLP} + \beta \mathcal{L}_{DET} \tag{3}$$

where β is the regularization factors. $\mathcal{L}_{DET}$ represents the general loss [30] in RPN and detection head.

During the training phase, the PLP, BPAP, PRN and detection head modules perform in an end-to-end fashion. During the inference phase, only RPN and detection head are carried out to obtain detection results, and both of PLP and BPAP are excluded, which bring litter overhead.

Fig. 2. Globally aware proposal generation. Green rectangle denotes the bounding-box proposal, and yellow one the discriminating part. Red point represents the matched predicted point. Red arrows and edges mean the shifting operation.

4 Experiments

4.1 Datasets

We conduct experiments on MSCOCO 2017 [25] datasets to evaluate the effectiveness of LPSNet. MSCOCO 2017 covers around 80 different object categories, including animals, everyday objects, and more. In experiments, we train LPSNet on MSCOCO 2017 *train* set ($115k$ images) and evaluate it on MSCOCO 2017 *test* set ($5k$ images).

4.2 Experiment Settings

All experiments are implemented on MMDetection [4]. We follow the default settings provided by MMdetection. The backbone is initialized with the ImageNet pre-trained weight. Input images are randomly resized while keeping the shorter sides do not exceed 800 and the longer sides do not exceed 1333. We choose AdamW as the optimizer with 0.05 weight decay a batch size of 16 in 8 GPUs and initialize the learning rate to 2e-4. We train the network for 12 epochs and decrease the learning rate at epochs 8 and 11 with a factor of 0.1.

4.3 Evaluation Metric

We use AP50, AP75, APs, APm and APl indicators to evaluate the accuracy of the proposrd network.

AP50: AP50 is the average precision at an IoU threshold of 50%. It assesses object detection accuracy with less stringent overlap criteria.

AP75: AP75 represents the average precision at an IoU threshold of 75%. It measures object detection accuracy with a stricter overlap threshold.

Method	Backbone	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
BS detectors						
RetinaNet [23]	ResNet-50	55.4	39.1	20.4	40.3	48.1
Reppoint [43]	ResNet-50	56.7	39.7	20.4	41.0	49.0
Faster R-CNN [30]	ResNet-50	58.1	40.4	21.2	41.0	48.1
Retrained PS detectors						
UFO [31]	VGG-16	27.9	-	-	-	-
UFO [31]	ResNet-50	28.9	-	-	-	-
P2BNet [5]	ViT-Small	43.5	13.6	7.6	19.1	31.3
End-to-end PS detectors						
Ours	ViT-Small	46.8	23.0	7.8	27.3	38.7
Ours*	ViT-Base	52.8	25.4	7.3	32.0	45.0

Table 1. The performance comparison of Box Supervised (BS), Retrained Point Supervised (PS), and End-to-end PS detectors on MSCOCO 2017 datasets. * indicates that the detection head part of LPSNet is appropriately modified based on imTED [47].

Fig. 3. Visualization of LPSNet results. Red bounding boxes represent predicted results, and green ones grounding truths.

APs, APm, APl: These metrics evaluate AP for small (APs), medium (APm), and large (APl) object instances separately. They focus on assessing object detection performance concerning objects of different scales, where small objects have an area less than 32×32, large objects have an area greater than 96×96, and medium objects fall in between.

4.4 Experimental Result

In experiment, we compare the LPSNet with the existing PSOD methods for comprehensive comparisons. In addition, to demonstrate the performance advantages of the PSOD methods, we compare the performance of the box-supervised object detectors to reflect their performance upper bound.

Comparison with Point-Supervised Methods. In the past, point supervised detectors adopted a retraining paradigm, which first generated box-level pseudo-labels and then used them to train the detector. In our proposed method, an end-to-end point supervised detector is designed to generate pseudo labels and detect objects of interest simultaneously. Experiments prove that the end-to-end paradigm achieves better performance. We compare LPSNet with state-of-the-art retrained point supervised detectors UFO^2 and P2BNet-FR. The results in Table 1 prove that LPSNet achieves significant performance improvements on various object scales and can more accurately locate object with high overlap rates.

Comparison with Box-Supervised Methods. Generally speaking, the accuracy of box-supervised detectors is much higher than that of point supervision. As shown in the last row of Table 1, benefiting from our proposed BLP and BPAP modules, LPSNet significantly reduces the accuracy gap between boxes and points labels.

Visualization of Results. To further verify the effectiveness of our method, we analyze it from a qualitative point of view. The red bounding boxes in Fig.3 show the prediction results of LPSNet, which are very close to the true value in green.

5 Conclusion

In this paper, we propose a point supervised detector that can adaptively regress local points to globally aware box-level proposals. LPSNet includes ViT-based backbone network, PLP, BPAP, RPN and detection head. Firstly, the backbone network takes concatenate patch tokens and points tokens as input to obtain valuable images and potential object features. Secondly, PLP confirms the location of the object at the point level. Third, BPAP adaptively regresses from points to component-level proposals and aggregating local parts into global proposals. Finally, under the supervision of the above global pseudo-proposals, LPSNet with RPN and detection head attached can be trained and inferred. Remarkably, the LPSNet takes full advantage of point information to generate high-quality proposals, which exhibits excellent detection accuracy under single point supervision.

References

1. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: IEEE CVPR. pp. 2846–2854 (2016)
2. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: IEEE CVPR. pp. 6154–6162 (2018)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229 (2020)
4. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R.,

Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
5. Chen, P., Yu, X., Han, X., Hassan, N., Wang, K., Li, J., Zhao, J., Shi, H., Han, Z., Ye, Q.: Point-to-box network for accurate object detection via single point supervision. arXiv preprint arXiv:2207.06827 (2022)
6. Chi, C., Wei, F., Hu, H.: Relationnet++: Bridging visual representations for object detection via transformer decoder. In: NeurIPS (2020)
7. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: IEEE CVPR. pp. 248–255. IEEE Computer Society (2009)
8. Deng, J., Pan, Y., Yao, T., Zhou, W., Li, H., Mei, T.: Relation distillation networks for video object detection. In: IEEE ICCV. pp. 7022–7031 (2019)
9. Diba, A., Sharma, V., Pazandeh, A., Pirsiavash, H., Van Gool, L.: Weakly supervised cascaded convolutional networks. In: IEEE CVPR. pp. 5131–5139 (2017)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
11. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. pp. 303–338 (2010)
12. Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J., Liu, W.: You only look at one sequence: Rethinking transformer in vision through object detection. arXiv preprint arXiv:2106.00666 (2021)
13. Fu, C., Liu, W., Ranga, A., Tyagi, A., Berg, A.C.: DSSD : Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659 (2017)
14. Gao, M., Li, A., Yu, R., Morariu, V.I., Davis, L.S.: C-wsl: Count-guided weakly supervised localization. In: ECCV (September 2018)
15. Gao, W., Wan, F., Pan, X., Peng, Z., Tian, Q., Han, Z., Zhou, B., Ye, Q.: TS-CAM: token semantic coupled attention map for weakly supervised object localization. In: IEEE ICCV. pp. 2886–2895 (2021)
16. Guo, Z., Liu, C., Zhang, X., Jiao, J., Ji, X., Ye, Q.: Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In: IEEE CVPR (June 2021)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR. pp. 770–778 (2016)
18. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: IEEE CVPR. pp. 3588–3597 (2018)
19. Jiang, P., Hou, Q., Cao, Y., Cheng, M., Wei, Y., Xiong, H.: Integral object mining via online attention accumulation. In: IEEE ICCV. pp. 2070–2079. IEEE (2019)
20. Kim, K., Lee, H.S.: Probabilistic anchor assignment with iou prediction for object detection. arXiv preprint arXiv:2007.08103 (2020)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1106–1114 (2012)
22. Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: IEEE CVPR. pp. 936–944 (2017)
23. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE ICCV. pp. 2999–3007 (2017)
24. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE ICCV. pp. 2999–3007 (2017)

25. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV. pp. 740–755 (2014)
26. 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)
27. Mingxiang Liao, Fang Wan, Y.Y.Z.H.J.Z.Y.W.B.F.P.Y.Q.Y.: End-to-end weakly supervised object detection with sparse proposal evolution. In: ECCV (2022)
28. Papadopoulos, D.P., Uijlings, J.R.R., Keller, F., Ferrari, V.: Training object class detectors with click supervision. In: IEEE CVPR. pp. 180–189 (2017)
29. Papadopoulos, D.P., Uijlings, J.R.R., Keller, F., Ferrari, V.: Training object class detectors with click supervision. In: IEEE CVPR (July 2017)
30. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NeuralIPS. pp. 91–99 (2015)
31. Ren, Z., Yu, Z., Yang, X., Liu, M., Schwing, A.G., Kautz, J.: Ufo^2: A unified framework towards omni-supervised object detection. In: ECCV. pp. 288–313 (2020)
32. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: IEEE ICCV. pp. 8429–8438 (2019)
33. Shen, Y., Ji, R., Chen, Z., Wu, Y., Huang, F.: UWSOD: toward fully-supervised-level capacity weakly supervised object detection. In: NeurIPS (2020)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
35. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., Luo, P.: Sparse R-CNN: end-to-end object detection with learnable proposals. In: IEEE CVPR. pp. 14454–14463 (2021)
36. Tang, P., Wang, X., Bai, S., Shen, W., Bai, X., Liu, W., Yuille, A.L.: PCL: proposal cluster learning for weakly supervised object detection. IEEE TPAMI pp. 176–191 (2020)
37. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: IEEE CVPR. pp. 3059–3067 (2017)
38. Tang, P., Wang, X., Wang, A., Yan, Y., Liu, W., Huang, J., Yuille, A.: Weakly supervised region proposal network and object detection. In: ECCV. pp. 352–368 (2018)
39. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. In: IEEE ICCV. pp. 9626–9635 (2019)
40. Wan, F., Wei, P., Han, Z., Jiao, J., Ye, Q.: Min-entropy latent model for weakly supervised object detection. IEEE Trans. Pattern Anal. Machine Intell. (10), 2395–2409 (2019)
41. Wang, C., Bochkovskiy, A., Liao, H.M.: Scaled-yolov4: Scaling cross stage partial network. In: IEEE CVPR. pp. 13029–13038 (2021)
42. Xie, S., Girshick, R.B., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: IEEE CVPR. pp. 5987–5995 (2017)
43. Yang, Z., Liu, S., Hu, H., Wang, L., Lin, S.: Reppoints: Point set representation for object detection. In: IEEE ICCV. pp. 9656–9665 (2019)
44. Zeng, Z., Liu, B., Fu, J., Chao, H., Zhang, L.: WSOD2: learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In: IEEE ICCV. pp. 8291–8299 (2019)
45. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: IEEE CVPR. pp. 4457–4465 (2017)

46. 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: IEEE CVPR. pp. 9756–9765 (2020)
47. Zhang, X., Liu, F., Peng, Z., Guo, Z., Wan, F., Ji, X., Ye, Q.: Integral migrating pre-trained transformer encoder-decoders for visual object detection. arXiv preprint arXiv:2205.09613 (2022)
48. Zhang, X., Wan, F., Liu, C., Ji, R., Ye, Q.: Freeanchor: Learning to match anchors for visual object detection. In: Neural Information Processing Systems. pp. 147–155 (2019)
49. Zhang, Y., Yan, Z., Sun, X., Diao, W., Fu, K., Wang, L.: Learning efficient and accurate detectors with dynamic knowledge distillation in remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–19 (2021)
50. Zhang, Y., Yan, Z., Sun, X., Lu, X., Li, J., Mao, Y., Wang, L.: Bridging the gap between cumbersome and light detectors via layer-calibration and task-disentangle distillation in remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–18 (2023)
51. Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: IEEE CVPR. pp. 2921–2929 (2016)
52. Zhu, C., He, Y., Savvides, M.: Feature selective anchor-free module for single-shot object detection. In: IEEE CVPR. pp. 840–849 (2019)
53. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: ICLR (2021)