PGTRNET: TWO-PHASE WEAKLY SUPERVISED OBJECT DETECTION WITH PSEUDO GROUND TRUTH REFINEMENT

Jun Wang[1], Hefeng Zhou[2],Xiaohan Yu[3]

[1]University of Warwick [2]Shanghai Jiao Tong University [3]Griffith University
jun.wang.3@warwick.ac.uk, zadeji1@sjtu.edu.cn, xiaohan.yu@griffith.edu.au

ABSTRACT

Current state-of-the-art weakly supervised object detection (WSOD) studies mainly follow a two-stage training strategy which integrates a fully supervised detector (FSD) with a pure WSOD model. There are two main problems hindering the performance of the two-phase WSOD approaches, i.e., insufficient learning problem and strict reliance between the FSD and the pseudo ground truth (PGT) generated by the WSOD model. This paper proposes pseudo ground truth refinement network (PGTRNet), a simple yet effective method without introducing any extra learnable parameters, to cope with these problems. PGTRNet utilizes multiple bounding boxes to establish the PGT, mitigating the insufficient learning problem. Besides, we propose a novel online PGT refinement approach to steadily improve the quality of PGT by fully taking advantage of the power of FSD during the second-phase training, decoupling the first and second-phase models. Elaborate experiments are conducted on the PASCAL VOC 2007 benchmark to verify the effectiveness of our methods. Experimental results demonstrate that PGTRNet boosts the backbone model by 2.1% mAP and achieves the state-of-the-art performance.

Index Terms— Weakly Supervised Object Detection, Pseudo Ground Truth, Online Refinement

1. INTRODUCTION

Object detection aims to detect and localize different semantic instances in images or videos with bounding boxes and obtain their categories. It is a significant sub-field of Computer Vision and has been extensively applied to many fields such as autonomous vehicles, security monitoring, and robotics. However, there are complex challenges since one image may contain different classes of objects and these instances may occur in different parts of images or videos.

Owing to the flourish of Deep Learning (DL) techniques, various DL-based algorithms have been proposed in object detection [1, 2, 3, 4]. Despite a great success has been achieved in these methods, they often require a large number of box-level annotations to improve their accuracy and generalization ability. This requirement, on the other hand, cannot always meet in real world applications. Therefore, Weakly Supervised Object Detection (WSOD), aiming to perform object detection with only image-level annotation, has gained increasing research attention. Most previous works [5, 6, 7, 8] exploit the Multiple Instance Learning (MIL) strategy for WSOD. Moreover, architectures combining the MIL and CNN demonstrate better performance due to the significant feature extraction ability of CNN compared with traditional hand-designed features. Research has shown that an end-to-end training or its variant can further improve the results for WSOD. Recent powerful WSOD methods normally follow a two-phase learning approach. A MIL detector combined with the CNN feature extractor are trained with image-level labels in the first phase. Outputs of the first-phase model are then taken as the pseudo ground truth (PGT) to fine-tune a Fully-Supervised Object Detection (FSOD) detector, e.g., Fast r-cnn [2].

However, most of previous works focus on the first-phase WSOD model and almost no study explores the potential of the second-phase training. We observe that current two-phase WSOD models regard the top one bounding box as the PGT, but the top one bounding box may only contain a part of instances or objects, and thus, may confuse the detectors. We denote this as the insufficient learning problem. In addition, PGT does not change during the whole training of the second-phase detector, hence even worse results will be obtained given the inaccurate PGT. To cope with these problems, we propose a pseudo ground truth refinement network (PGTRNet) which takes advantage of multiple bounding boxes to produce the PGT. Furthermore, PGTRNet gradually refines the PGT to release the reliance of the WSOD detector.

To sum up, this work mainly has four contributions. **(1)** We propose a novel two-phase WSOD architecture, the PGTRNet, which takes advantage of PGT by the first-phase detector training and gradually refines them during the second-phase training, alleviating the reliance between the WSOD and FSOD models. Moreover, PGTRNet is simple yet effective without introducing any extra learnable parameters. **(2)** Different from previous works that only regard the top one bounding box as PGT, we propose to consider more bounding boxes to generate the PGT, mitigating the insufficient learning problem. **(3)** We develop multiple strategies to effectively refine the PGT during the training of the second-phase detector. **(4)** We verify the effectiveness of our PGTRNet on the VOC2007 benchmark. Experimental results demonstrates that PGTRNet improves the backbone network by 2.1% mAP and achieves state-of-the-art performance. Further ablation studies show the necessity of PGT refinement in the FSOD detector training.

2. RELATED WORK

The main issue of WSOD is that it is much harder to generate the bounding box of objects as no ground truth location is available. In WSOD, the model may be misled and this may be amplified if the given supervision is always inaccurate. To solve the WSOD problem, current methods normally adopt the Multiple Instance Learning (MIL) [9] strategy and take advantage of CNN, and form an end-to-end unified learning approach.

Weakly Supervised Deep Detection Network (WSDDN), proposed by Bilen & Vedaldi [5] is a two-stream architecture. One stream for classification and one stream for detection. The positive sample is determined in terms of the score of these streams. Based on WSDDN, Kantorov et al. [10] introduces the use of context (ei-

ther additive or contrastive) to aid in localization. This is also in accordance with the spirit of recent research [11, 12, 13, 14], which focuses on localizing vital regions. One major problem of these methods is that the algorithm tends to get trapped in a local minimum.. To alleviate this problem, Tang et al. [15] presents an online instance classifier refinement (OICR) algorithm which is also based on WSDDN. Also proposed by Tang et.al. [16], Proposal Cluster Learning (PCL) clusters the region proposals of an image into different clusters before sending them as the inputs to the backbone network, making the algorithm pay attention to the whole part of object.

Recent state-of-the-art WSOD approaches follow a two-phase learning strategy. A weakly supervised network (e.g., methods mentioned above) is first trained to obtain the predicted bounding boxes on a training dataset. These predicted bounding boxes are then regarded as the PGT to fine-tune a fully supervised detector. Most approaches [16, 15, 17, 18] exploit Fast r-cnn as the second-phase network and outperform the pure WSOD model. However, as mentioned in the previous section, insufficient learning problem limits the performance of the these methods. Besides, the effectiveness of the fully-supervised detector is highly dependent on the quality of the PGT generated by the WSOD detector.

3. METHODS

This section describes our proposed methods to advance the precision of WSOD. PGTRNet adopts the PCL [16] and the Faster r-cnn [19] as the backbone networks of the first-phase and second-phase detectors, respectively. We first briefly review the general framework of the two-phase WSOD model. Our PGTRNet can be built on top of any two-phase WSOD architecture. Figure 1 illustrates the general architecture of PGTRNet. Images and the associated region proposals are fed into the first-phase WSOD detector to train the detector and obtain the proposal scores S_c. PGT generation is then applied to produce the PGT according to the region proposals and their corresponding scores. Combined with the images, PGT is then sent to the fully supervised detector (FSD) to supervise the learning. During the training of the FSD, we suspend the training and apply the PGT refinement multiple times to improve the quality of PGT based on the refinement strategy. The details of PGTRNet including PGT generation and PGT refinement are described in section 3.2.

3.1. Two-phase WSOD

We denote a training image as $I \subseteq R^{H \times W \times 3}$ and its image-level annotation as $y = [y_1, y_2, ..., y_C] \subseteq [0, 1]^C$. Note that H and W are the height and width of the image. C is the total number of classes. Given the I and y, WSOD aims to predict the potential objects (bounding boxes) B_c for class c and corresponding scores S_c. N is the total number of the outputs from the WSOD model.

$$B_c = \{b_{c,1}, b_{c,2}, ..., b_{c,N}\} \tag{1}$$

$$S_c = \{s_{c,1}, s_{c,2} ..., s_{c,N}\} \tag{2}$$

Current state-of-the-art WSOD methods often follow a two-phase training strategy. Specifically, in the first stage, a pure WSOD model is trained with only I and y. Then, training images I are fed into the trained WSOD model again to obtain the predicted bounding boxes B_c for class c. After obtaining B_c, top one bounding box b_c^* is selected as the pseudo ground truth to fine-tune a fully-supervised

detector. Equations (3) shows this procedure.

$$k = argmax_i(s_{c,i}), \quad b_c^* = b_{c,k} \tag{3}$$

3.2. PGTRNet

3.2.1. PGT Generation

PGT is the core to connect the first-phase and second-phase models. As noted above, most previous works select the top one bounding boxes as the PGT to train a second-phase model such as Fast r-cnn. It is very natural to select the highest confidence bounding box since detectors tend to be trained with correct supervision according to these high score boxes. Nonetheless, this could lead to some problems hindering the performance of the detector. We empirically observe that objects from the same class may appear more than once in a large number of pictures and datasets. In this case, only given the top one bounding boxes, the detector may be unable to learn adequate features to distinguish this class from others. This problem could be exacerbated when the dataset is small or there are many subcategories in a class such as Border Collie and Akita in the dog category. Besides, it is likely that the top bounding boxes are misclassified or contain only parts of an object in WSOD. These cases may also cause an insufficient learning problem.

To mitigate the problems mentioned above, we propose to consider top k bounding boxes as the PGT to guarantee a high proposal recall. Nonetheless, a large k value may also introduce many negative samples, and thus mislead the learning while a small k value may lose significant learning information. Hence, it is essential to maintain the balance between high proposal recall and high precision. We collect top k bounding boxes in three steps.

(1) We train a WSOD model according to the images and associated image-level annotation, and then perform inference on the same dataset to obtain the predicted bounding boxes and their scores.

(2) Recall that algorithms do not know the number of instances in each category in images and their bounding boxes but they do have knowledge about which class occurs in the image for WSOD. We record this information via processing the annotation file.

(3) Based on information obtained from step one and two, top k bounding boxes B_c' for class c are created by ordering the bounding boxes based on their scores S_c in corresponding categories and selecting the best k boxes. We then save the PGT to establish an instance-level annotation file which will be processed to train the second-phase model. Equations (3) then changes to be Equations (4) and (5).

$$L_c = argsort(B_c)[1 \sim k] \tag{4}$$

$$B_c' = \{b_{c,1}, b_{c,2}, .., b_{c,i}, .., b_{c,k}\}, i \subseteq L_c \tag{5}$$

3.2.2. PGT Refinement

Previous works often focus on designing decent pure WSOD model and neglect the second-phase fine-tuning procedure. We argue that the second-phase training has huge potential to improve the results of the WSOD model. Given the PGT, it has been shown that a fully supervised model may improve the results from a pure WSOD model. However, for those misclassified samples and samples only containing non-distinguishable parts, a second-phase detector seems powerless. Considering this, we propose to refine the PGT during the training. This strategy can improve the PGT or increase the number of positive training samples, and thus may boost the performance.

PGTRNet performs inference on the training dataset during the training of the fully supervised detector and recreates instance-level
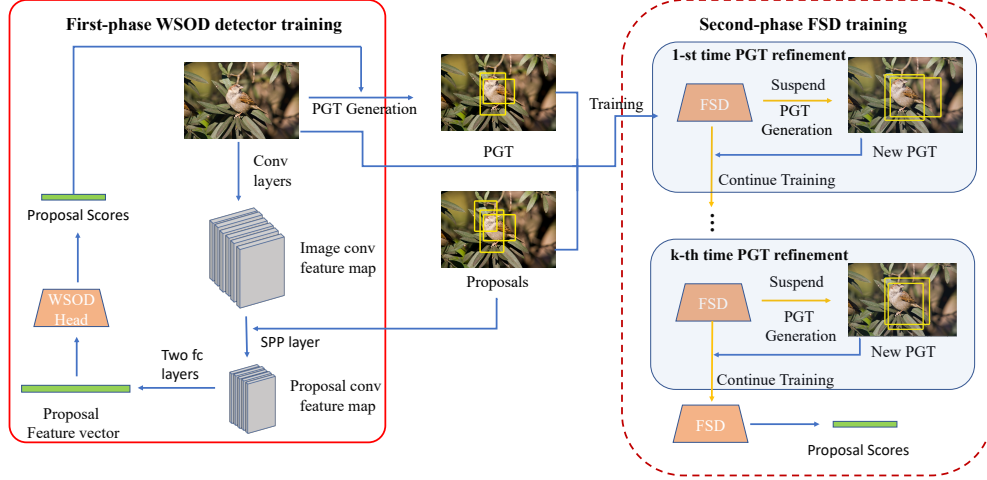
Fig. 1 – The general architecture of the PGTRNet. We first train the first-phase WSOD detector according to the training images and the corresponding region proposals. Inference is then applied to these images to obtain the proposal scores. PGTRNet then produces the PGT via the proposed PGT generation method. The PGT is then utilized to supervised the learning of the fully supervised detector (FSD). To improve the quality of the PGT and alleviate the reliance between the first and second-phase detector, PGTRNet refines the PGT multiple times based on the outputs of the FSD and the refinement strategies.

annotations according to the bounding box scores S_c. Unlike the PGT generation in Section 3.2.1, this update is performed online. This means that the weights and gradients of the network are maintained after performing inference on training dataset. There are two reasons for choosing this strategy. **(1)** Compared with completing training and then refining the PGT in a new model, this strategy has lower computational complexity. **(2)** During the training, the detector may not be misled from those negative training samples much.

Then, the problem turns to how and when to refine the PGT. We propose seven strategies to refine the PGT. These strategies are divided into two groups as shown in Table 1 Strategies in group one determine the time to refine the PGT, while those in group two define how to update the PGT. A complete refinement strategy comprises one strategy from group one and one strategy from group two.

To be specific, we do PGT refinement in the following steps. As a necessary procedure, a PCL model is trained to generate the annotations (PGT) for the second-phase detector via the method described in Section 3.2.1. We then train the second-phase model based on the PGT. At specific epochs during the training (e.g., epoch mentioned in group one), we suspend training, save the model, and change the model from training mode to evaluation mode. After suspending the training, we perform inference on training images and then update the annotation files according to the strategies mentioned in group two. One thing that should be emphasized is that we do not stop the training, and thus, the gradient and the weight of the network are kept. The next step is to continue training according to the refined annotations.

4. EXPERIMENTS

We verify the effectiveness of PGTRNet on one of the most common WSOD benchmark PASCOL VOC 2007 dataset. Mean Average Precision (mAP) is used to evaluate the model. Section 4.1 shows the details to reproduce our work. Experimental results and comparisons to SOTA are given in the following sections. Eventually, an ablation study is demonstrated to further prove the effectiveness of

Table 1 – Group one and two of the PGT refinement strategies.

Group 1:
1. Refine the PGT after each epoch.
2. Refine the PGT every three epochs.
3. Refine the PGT after each epoch in the last three epochs.
4. Refine PGT only once at the 2/3 of the maximum epochs.

Group 2:
1. Update all k PGT .
2. Only update half of the k best PGT.
3. Only update half of the k worst PGT.

our propose methods.

4.1. Implement Details & Dataset

Following the most previous two-phase works, we use five image scales {480, 576, 688, 864, 1200} to enable multi-scale training and testing via resizing the shortest side to the closest scales. k in equations (4) and (5) is set to 2. We train all models with 12K iterations and the minibatch size is set to 4. The initial learning rate is set to 0.002 and decayed in 8000 and 10500 iterations. All experiments are conducted on one Nvidia GTX 2080Ti GPU card.

The PASCAL VOC 2007 [26] dataset is one of the most common benchmarks to measure the ability of a model in object detection. A large amount of works test their models and report the results on this dataset. This benchmark consists of 5011 training images and 4952 testing images with a total amount of 20 categories.

4.2. Detection Results

Here, we demonstrate the detection results of PGTRNet and comparisons to state-of-the-art methods. Table 2 gives comparisons of PGTRNet to recent state-of-the-art methods on PASCAL VOC 2007 dataset. Obviously, PGTRNet beats all methods with a mAP of 55.3% and outperforms the second best-performed method by 0.4%. Moreover, PGTRNet achieves the best average precision on almost

Table 2 – The comparisons to state-of-the-art methods on the test set of PASCAL VOC 2007. The last line is our proposed method. The best AP of each class and mAP are highlighted.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mobile	person	plant	sheep	sofa	train	tv	mAP
PCL-Ens+FRCNN [16]	63.2	69.9	47.9	22.6	27.3	71.0	69.1	49.6	12.0	60.1	51.5	37.3	63.3	63.9	15.8	23.6	48.8	55.3	61.2	62.1	48.8
ML-LocNet-L+ [10]	60.8	70.6	47.8	30.2	24.8	64.9	68.4	57.9	11.0	51.3	55.5	48.1	68.7	69.5	28.3	25.2	51.3	56.5	60.0	43.3	49.7
WSRPN-Ens+FRCNN [18]	63.0	69.7	40.8	11.6	27.7	70.5	74.1	58.5	10.0	66.7	**60.6**	34.7	**75.7**	70.3	25.7	26.5	55.4	56.4	55.5	54.9	50.4
Multi-Evidence [20]	64.3	68.0	56.2	36.4	23.1	68.5	67.2	64.9	7.1	54.1	47.0	57.0	69.3	65.4	20.8	23.2	50.7	59.6	65.2	57.0	51.2
W2F+RPN+FSD2 [21]	63.5	70.1	50.5	31.9	14.4	72.0	67.8	73.7	23.3	53.4	49.4	65.9	57.2	67.2	27.6	23.8	51.8	58.7	64.0	62.3	52.4
OIM+IR+FRCNN [22]	53.4	72.0	51.4	26.0	27.7	69.8	69.7	74.8	21.4	67.1	45.7	63.7	63.7	67.4	10.9	25.3	53.5	60.4	70.8	58.1	52.6
PredNet-Ens [23]	**67.7**	70.4	52.9	31.3	26.1	**75.5**	73.7	68.6	14.9	54.0	47.3	53.7	70.8	70.2	19.7	29.2	54.9	**61.3**	67.6	61.2	53.6
WSOD2 [24]	65.1	64.8	57.2	**39.2**	24.3	69.8	66.2	61.0	**29.8**	64.6	42.5	60.1	71.2	70.7	21.9	28.1	58.6	59.7	52.2	64.8	53.6
GAM+PCL+Reg [25]	59.8	72.8	54.4	35.6	30.2	74.4	70.6	**74.5**	27.7	68.0	51.7	46.3	63.7	68.6	14.8	27.8	54.9	60.9	65.1	67.4	54.5
Wetectron [4]	68.8	**77.7**	57.0	27.7	28.9	69.1	**74.5**	67.0	32.1	**73.2**	48.1	45.2	54.4	**73.7**	**35.0**	29.3	**64.1**	53.8	65.3	65.2	54.9
PGTRNet (ours)	65.6	73.0	**61.2**	29.9	**34.4**	69.0	67.9	60.7	28.8	67.7	52.9	**71.6**	68.5	67.2	12.3	**30.9**	43.2	59.1	**72.1**	**70.0**	**55.3**

Table 3 – The mAP of different combination of strategies of PGTR-Net on the test set of PASCAL VOC 2007 benchmark. − (no strategy been utilized, or no results). The best mAP is highlighted in bold.

mAP GP2 / GP1	−	1	2	3
−	53.54	−	−	−
1	−	27.85	55.14	54.44
2	−	45.88	54.33	**55.29**
3	−	45.31	53.78	54.59
4	−	54.56	54.00	53.87

a third of all categories (6 out of 20). These encouraging results confirm the superiority of PGTRNet. Table 3 shows the performance of PGTRNet with different combination of PGT refinement strategies from groups one and two. We denote a complete strategy as */* and * is the $*^{th}$ strategy of group1/2. Note that the proposed PGT generation is added on all experiments including the group −/−. It can be observed that strategies 2 and 3 of group two combining with any strategy from the group one boost the performance by almost 1% mAP. These strategies only updates half of the PGT, preventing the network from significant fluctuation while refining the PGT. The best strategy 1/2 improves the mAP from 53.54% to 55.29% (1.74% mAP) which achieves the state-of-the-art performance.

However, strategies 1,2,3/1 noticeably limit the performance possibly due to the fluctuation of the network. Strategy 1 of group two updates all the PGT, and the model fails to converge if the algorithm frequently performs refinement. A different result can be seen in strategy 4/1 which boosts the mAP by 1%. A possible reason is that the algorithm performs refinement only once in strategy 4 of group one, and the detector still has adequate training epochs to fit the updated PGT. By comparing the results of strategy 1/2, 1/3, 2/2, 2/3, we can conclude that it is detrimental to perform the PGT refinement too frequently.

Table 4 – Ablation studies of the hyper-parameter k. The best accuracy is highlighted in bold.

k	1	2	3	4	5
mAP	53.21	53.53	**53.99**	53.63	53.21

4.3. Ablation study

We conducted an ablation study to evaluate the effectiveness of each component of PGTRNet. We first investigate the influence of different settings of k on the PGT generation. The effectiveness of

Table 5 – The ablation study results of each component (PGT-G and PGT-R) of PGTRNet on the test set of PASCAL VOC 2007 benchmark.

PGT-G	PGT-R	mAP
-	-	53.21
√	-	53.99
√	√	55.29

considering more bounding boxes on PGT generation (PGT-G) and PGT refinement (PGT-R) are then discussed. Note that experiments on exploring the influence of the value of k is not combined with the proposed PGT refinement.

As shown in Table 4, the mAP increases modestly from 53.21% to 53.99% as k changes from 1 to 3. The main reason for this improvement is that an image may contain multiple instances from the same categories and a larger k value allows algorithms to consider this case. This confirms the necessity of considering more bounding boxes when generating the PGT. Nonetheless, a different trend can be seen when k continues to increase. The mAP reduces modestly from 53.99% to 53.21% when k increases to 5. These reductions are caused by introducing more negative samples in PGT than positive samples.

Table 5 demonstrates the effectiveness of each component (PGT-G and PGT-R) of PGTRNet. PGT-G denotes the method of taking advantage of top three bounding boxes when generating the PGT. PGT-R is the PGT refinement strategy 2/3 with top two bounding boxes since we should ensure that the half of the k is an integer on PGT refinement. It can be seen that PGT-G modestly boosts the mAP from 53.21% to 53.99%. PGT-R further improves the mAP by 1.30% to 55.29%. Combined with PGT-G, a significant improvement can be seen on PGT-R from 53.21% to 55.29% (+2.08%) which achieves the state-of-the-art performance.

5. CONCLUSION

Recent state-of-the-art WSOD approaches integrate the WSOD detector and fully-supervised detector. These methods often neglect the significance of the second-phase training, suffering from the insufficient learning problem. We propose to consider more bounding boxes when creating the PGT. The low quality PGT generated by the WSOD detector deleteriously affects the performance of the fully supervised detector. To mitigate these problems, we propose a novel two-phase WSOD model PGTRNet which takes advantage of multiple bounding boxes to generate the PGT and gradually improves the quality of PGT. PGTRNet is simple yet effective without introducing extra learnable parameters. Experimental results and the ablation study demonstrate the superiority of PGTRNet which achieves the state-of-the-art performance on PASCAL VOC 2007 benchmark.

6. REFERENCES

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[2] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[3] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[4] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz, "Instance-aware, context-focused, and memory-efficient weakly supervised object detection," in *Proceedings of the IEEE Conference on Computer Bision and Pattern Recognition*, 2020, pp. 10598–10607.

[5] Hakan Bilen and Andrea Vedaldi, "Weakly supervised deep detection networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2846–2854.

[6] Xinggang Wang, Zhuotun Zhu, Cong Yao, and Xiang Bai, "Relaxed multiple-instance svm with application to object discovery," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1224–1232.

[7] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid, "Weakly supervised object localization with multifold multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 189–203, 2016.

[8] Peng Tang, Xinggang Wang, Zilong Huang, Xiang Bai, and Wenyu Liu, "Deep patch learning for weakly supervised object classification and discovery," *Pattern Recognition*, vol. 71, pp. 446–459, 2017.

[9] Alessandro Prest, Cordelia Schmid, and Vittorio Ferrari, "Weakly supervised learning of interactions between humans and objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 601–614, 2011.

[10] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev, "Contextlocnet: Context-aware deep network models for weakly supervised localization," in *European Conference on Computer Vision*. Springer, 2016, pp. 350–365.

[11] Jun Wang, Xiaohan Yu, and Yongsheng Gao, "Mask guided attention for fine-grained patchy image classification," in *28th IEEE International Conference on Image Processing*, 2021.

[12] Xiaohan Yu, Yang Zhao, Yongsheng Gao, Shengwu Xiong, and Xiaohui Yuan, "Benchmark platform for ultra-fine-grained visual categorization beyond human performance," in *International Conference on Computer Vision (ICCV)*, 2021.

[13] Xiaohan Yu, Yang Zhao, Yongsheng Gao, and Shengwu Xiong, "Maskcov: A random mask covariance network for ultra-fine-grained visual categorization," *Pattern Recognition*, p. 108067, 2021.

[14] Xiaohan Yu, Yang Zhao, Yongsheng Gao, Shengwu Xiong, and Xiaohui Yuan, "Patchy image structure classification using multi-orientation region transform.," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2020, pp. 12741–12748.

[15] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu, "Multiple instance detection network with online instance classifier refinement," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2843–2851.

[16] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille, "Pcl: Proposal cluster learning for weakly supervised object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 1, pp. 176–191, 2018.

[17] Xiaopeng Zhang, Yang Yang, and Jiashi Feng, "Ml-locnet: Improving object localization with multi-view learning network," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 240–255.

[18] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille, "Weakly supervised region proposal network and object detection," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 352–368.

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99, 2015.

[20] Weifeng Ge, Sibei Yang, and Yizhou Yu, "Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1277–1286.

[21] Yongqiang Zhang, Yancheng Bai, Mingli Ding, Yongqiang Li, and Bernard Ghanem, "W2f: A weakly-supervised to fully-supervised framework for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 928–936.

[22] Chenhao Lin, Siwen Wang, Dongqi Xu, Yu Lu, and Wayne Zhang, "Object instance mining for weakly supervised object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 11482–11489.

[23] Aditya Arun, CV Jawahar, and M Pawan Kumar, "Dissimilarity coefficient based weakly supervised object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9432–9441.

[24] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang, "Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8292–8300.

[25] Ke Yang, Dongsheng Li, and Yong Dou, "Towards precise end-to-end weakly supervised object detection network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8372–8381.

[26] Mark Everingham, L Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes challenge 2007 (voc 2007) results (2007)," 2008.