Proposal Feature Learning Using Proposal Relations for Weakly Supervised Object Detection

Zhaofei Wang [1,2], **Weijia Zhang** [3], **Min-Ling Zhang** [1,2*]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China
[3]School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia
wangzf@seu.edu.cn, weijia.zhang@newcastle.edu.au, zhangml@seu.edu.cn

***Abstract*—Weakly Supervised Object Detection (WSOD) trains detectors using only image-level annotations. Most existing WSOD models are based on pre-computed proposals and do not fully explore the relations of proposals. In this work, we address this limitation by proposing two approaches of Proposal Feature Learning for WSOD (PFL-WSOD), which effectively capture *intra-proposal* relations and *inter-proposal* relations respectively, thus improving proposal representation. To extract intra-proposal relations, we propose to utilize Self-Attention on Single Proposal for capturing relations inside each proposal. For inter-proposal relations, we propose Salient Region Banks by capturing a unique type of inter-proposal relation called *deep inclusion*, which significantly improves proposal representation when used in synergy with contrastive learning. Experimental results on benchmarks demonstrate the effectiveness of our methods.**

***Index Terms*—Weakly supervised object detection, Self-attention, Proposal relations**

I. Introduction

Over the last decade, remarkable results have been achieved by various object detection algorithms such as Fast-RCNN [1] in Fully Supervised Object Detection (FSOD) tasks. This success is largely attributed to the availability of large datasets with accurate object categories and meticulously annotated bounding boxes, which is laborious, time-consuming, and typically incurs substantial costs.

To address this challenge, Weakly Supervised Object Detection (WSOD) has emerged as an alternative approach, utilizing only image-level annotations and omitting the need for bounding box locations in the training phase. The significantly reduced annotation costs associated with WSOD have led to a notable surge in research interest in this domain over recent years, as highlighted by studies such as [2]–[5].

However, the effectiveness of existing WSOD algorithms still lags behind that of FSOD methods, primarily due to weaker supervision signals and algorithmic shortcomings. The limitations in the performance of existing WSOD algorithms stem from two main issues: the *salient region* problem [4] and the *missing instance* problem [4]. The former issue misleads the model to prioritize the most distinctive features of the image, e.g., the head of the bird contained in the red box in Figure 1, instead of the complete object. The latter issue occurs when multiple instances of the same class are present, yet the detector fails to identify all of them.

*Corresponding author.

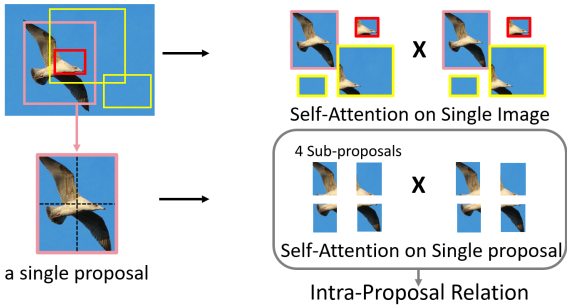

Fig. 1. We propose the self-attention mechanism on each proposal for capturing intra-proposal relationships. Unlike self-attention on a single image (first row), our method captures the nuances of intra-proposal relationships by focusing on self-attention on each single proposal (second row).

Despite various attempts to address these challenges, the potential of leveraging proposal relations has not been fully explored. Although self-attention (SA) [6] has seen application in WSOD [7], [8], they abandon pre-computed proposals and utilize SA as an image-level feature extractor and a bounding box generator following DETR [9], instead of exploring proposal relations. In this work, we argue that utilizing proposal relations is essential for learning better proposal features in WSOD. To this end, we introduce two approaches of Proposal Feature Learning for WSOD (PFL-WSOD) designed for capturing two distinct types of proposal relations: *intra-proposal* relations and *inter-proposal* relations. It is worth noting that, our PFL is a kind of self-supervised learning, different from updating proposal features by supervisions of image or proposal classification.

Our *intra-proposal* relations pertain to the relationships among various segments of a proposal, i.e., sub-proposals, as shown in the second row of Figure 1. We introduce PFL-WSOD_$intra$, which utilizes self-attention on single proposal to capture this type of relations. Thanks to the significantly reduced size and number of sub-proposals, PFL-WSOD_$intra$ is able to efficiently discern relations among sub-proposals and thus improves feature representation for each proposal.

Our *inter-proposal* relations concern the connections among different proposals. To this end, we propose PFL-WSOD_$inter$ which establishes a *Salient Region Bank* for each class to gather proposals of salient regions by utilizing

a special type of *inter-proposal* relation 'deep inclusion' in different refinement branches. If one proposal is greatly contained within another proposal spatially, we say that these two proposals have 'deep inclusion' relation. Contrastive learning is adopted to increase the distance between training proposals and proposals in *Salient Region Banks*, finally realizing PFL.

Learning better proposal features using PFL-WSOD leads to better detection results, despite the lack of instance-level supervision. Experiments on benchmarks report better results compared with state-of-the-arts. Specifically, we achieve 60.4% in mAP on VOC2012 test set, which is significantly higher than state-of-the-arts. We also achieve highest and comparable performances on VOC2007 and COCO.

II. The Proposed Method

A. Preliminaries

We first introduce our baseline WSOD network. An input image accompanied by its label $Y = \{y_1, ..., y_C\}$ without bounding boxes during training and a total of N candidate proposals pre-computed by external method are fed into a CNN feature extractor. The extracted features after ROI Pooling layers is flattened. We denote the flattened features as proposal feature vectors $P = \{\boldsymbol{p}_1, ..., \boldsymbol{p}_N\}$ where $\boldsymbol{p}_n \in \mathbb{R}^D$ and their corresponding candidate bounding boxes as $B = \{\boldsymbol{b}_1, ..., \boldsymbol{b}_N\}$. Then, P is sent to the Multi-Instance Learning (MIL) module [3] [10] [11] to generate proposal scores $X \in \mathbb{R}^{C \times N}$. Finally, $\phi_c = \sum_{n=1}^{N} x_{cn}$ is aggregated as the image-level prediction, which is used to calculate the cross entropy loss $\mathcal{L}_{CE} = -\sum_{c=1}^{C} y_c \log \phi_c + (1 - y_c) \log(1 - \phi_c)$. Online Instance Refinement (OIR) [3] is utilized following the MIL module to alleviate the *salient region* problem. The refinement process consists of K branches, where each branch utilizes the proposal features as input and produces a proposal classification score matrix X^k. Next, supervision for the classification scores of each branch is determined by labeling proposals having high IoU (Intersection over Union) with the top1-scored proposal from the previous branch as positive instances. As a result, the k-th branch is a refinement for the $k-1$-th branch. OIR is trained with these supervisions by a proposal classification loss function $\mathcal{L}_{OIR}$. Also, bounding box regression is adopted after the last refinement branch.

B. PFL-WSOD

The key idea of our proposed PFL-WSOD approaches is to facilitate better proposal feature learning by exploiting proposal relations, a method akin to self-supervised learning. This approach is fundamentally different from the traditional focus on enhancing supervision for refinement branches as commonly seen in previous WSOD methods.

Better proposal features can improve detection performances under weak supervision. Let us consider global region proposals as proposals containing the complete object and salient region proposals as proposals containing only salient regions. As depicted in Figure 3, while the top1-scored proposal selected by the MIL module in early training iterations is desirable, it eventually degenerates to salient regions.

The performance degradation occurs because the similarity score between salient and global region proposals, which we term Similarity$(salient, global)$[1], decreases during early or middle stages of training but subsequently exhibits a noticeable increase as depicted in Figure 4. The high similarity misleads the detector to perceive these two types of proposals as similar, consequently leading to inaccurate detection. This issue is illustrated in the first row of Figure 5.

Therefore, as shown in the second row in Figure 5, our PFL-WSOD is designed to ensure a decreasing trend in Similarity$(salient, global)$. By doing so, we ensure that these two types of proposals remain sufficiently distinct during the subsequent training phases, ultimately enhancing the overall detection performance. Since proposal features are shared by both the MIL classification branch and refinement branches, improved proposal features will relieve both *salient region* problem and *missing instance* problem. The architectures of our methods are illustrated in Figure 2.

C. PFL-WSOD with Intra-Proposal Relations

Self-attention [6] has been proven to be effective for extracting relationships among tokens in natural language processing tasks [12]. A naive implementation of self-attention in WSOD would be to utilize a multi-layer self-attention architecture, for learning the relationships among proposals of different sizes generated from a single image (as shown in the first row of Figure 1). However, this naive approach will lead to excessive time and memory complexities as it involves calculating a matrix multiplication $A^T B$ ($A, B \in \mathbb{R}^{D \times N}$ with $D = 4096$ and $N \approx 2000$) for multiple self-attention layers. Also, capturing relations among such long, high-dimensional sequences with weak supervision is not easy.

Therefore, to fully leverage the capabilities of the self-attention mechanism, we propose to explore relations within each proposal and define *intra-proposal* relations as the relations among small parts of a certain proposal. The *self-attention on single proposal* depicted in the second row of Figure 1 illustrates our proposed approach. Given a proposal $\boldsymbol{p}_n \in \mathbb{R}^D$, we split it into M sub-proposals $\boldsymbol{p}'_n = \{\boldsymbol{p}^{sub}_{n1}, ..., \boldsymbol{p}^{sub}_{nM}\} \in \mathbb{R}^{D' \times M}$ with each sub-proposal $\boldsymbol{p}^{sub}_{nm} \in \mathbb{R}^{D'}$, where $M \times D' = D$. Equation (1) - (4) summarize the procedure of our proposed *self-attention on single proposal*, which is also illustrated in Figure 2. W_1, W_2 and W_3 are learnable parameters. Additionally, we incorporate residual blocks and feed-forward networks, following Vaswani *et al.* [6].

$$q_n, k_n, v_n = W_1 \boldsymbol{p}'_n, \quad W_2 \boldsymbol{p}'_n, \quad W_3 \boldsymbol{p}'_n \tag{1}$$

$$q_n, k_n, v_n \in \mathbb{R}^{D' \times M} \tag{2}$$

$$A_n = \frac{q_n^T k_n}{\sqrt{D'}} \in \mathbb{R}^{M \times M} \tag{3}$$

$$\boldsymbol{p}''_n = v_n \cdot \text{Softmax}(A_n) \tag{4}$$

The learned representations $\boldsymbol{p}''_n$ are then reshaped to $\mathbb{R}^D$ and collected as self-attention proposal features (SA features

[1]Details of calculating similarities are available in the experiments.

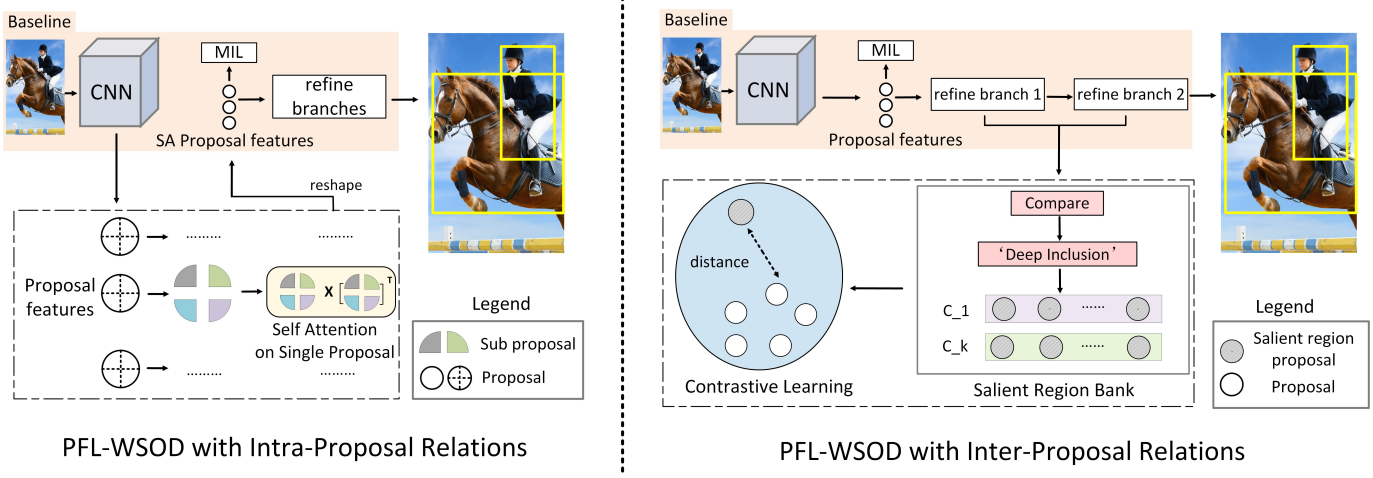

Fig. 2. Overview of PFL-WSOD architecture. We insert our methods into baseline. PFL-WSOD_$intra$ learns relationships among sub-proposals using *Self-Attention on Single Proposal*. PFL-WSOD_$inter$ establishes *Salient Region Banks* by determining 'deep inclusion' relation between proposals. Contrastive learning is implemented based on *Salient Region Banks*.

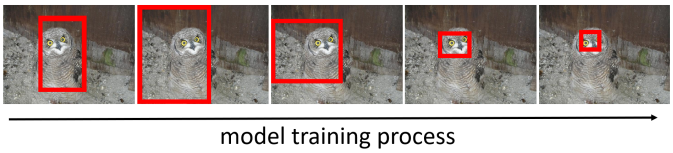

Fig. 3. Without PFL, the model detects global region proposals (left) in early epochs but converges to salient region proposals (right) as the training progresses in some cases.

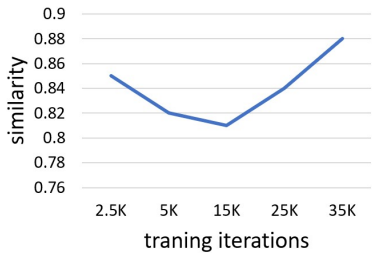

Fig. 4. The similarity between salient region proposals and global region proposals of the training image in Figure 3.

in Figure 2), which are used as the input for the MIL branch and the refinement branches for alleviating both the *salient region* and the *missing instance* problems. Experiments illustrate that PFL-WSOD_$intra$ effectively decreases Similarity$(salient, global)$.

D. PFL-WSOD with Inter-Proposal Relations

The self-attention-based PFL-WSOD_$intra$ is an unsupervised approach. Alternatively, we can design and extract more specific types of proposal relations, followed by implementing carefully crafted rules to refine the proposal features. In this section, our focus shifts to exploring a specific type of inter-proposal relations with PFL-WSOOD_$inter$.

Wang *et al.* [5] explored the concept of using contrastive learning for encouraging proposals to distance themselves

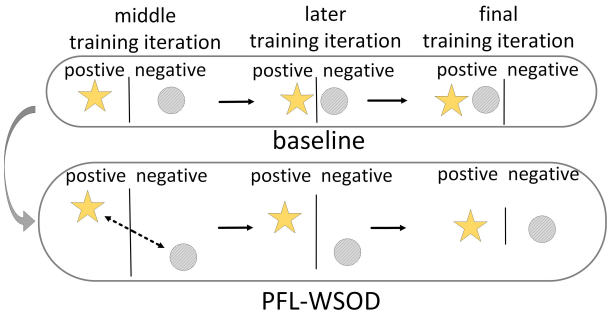

Fig. 5. The comparison between baseline and PFL-WSOD. Yellow stars represent global region proposals, and grey circles represent salient region proposals.

from incorrect instances. Given a image of class c, when its proposal $\boldsymbol{p}_n^c$ is misclassified to class $c' \neq c$, they treat $\boldsymbol{p}_n^c$ (incorrect instances) as a salient region of c'. However, their approach leads to cross-class approximation for the salient regions, which will introduce bias as proposals of other classes are inappropriately identified as salient regions of class c'.

To mitigate the bias in identifying the incorrect instances of class c', we propose an intuitive approach by leveraging *inter-proposal* relations. Different from the previous approach [5], our approach regards proposals that only contains the salient regions of class c' as incorrect instances of class c'. When combined with contrastive learning, this approach naturally reduces Similarity$(salient, global)$.

Identifying the salient regions is challenging. A naive approach for finding the salient regions is to regard the top1-scored proposal of the MIL branch or the first refinement branch as salient region proposals as they often fall into local optima. However, as the training iteration increases, these proposals may escape from local optima and are no longer salient region proposals. To address this challenge, we

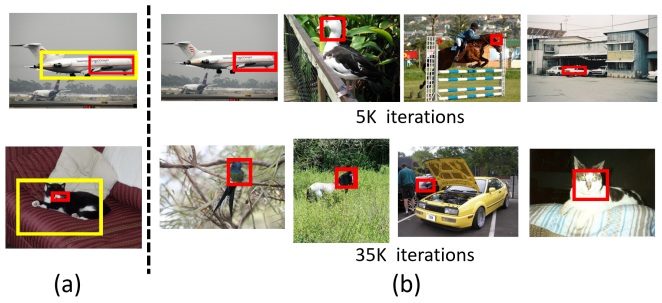

Fig. 6. (a) Explanation of 'deep inclusion' relation. The red and yellow box indicate the top1-scored proposal in the first and the last refinement respectively. (b) Salient region proposals (red boxes) selected randomly from different Salient Region Banks in different training iterations.

define a special type of *inter-proposal* relation called 'deep inclusion'. Given two bounding boxes $\boldsymbol{b}_i, \boldsymbol{b}_j \in B$ from an image, $deep\_inclusion(\boldsymbol{b}_i, \boldsymbol{b}_j)$ returns $True$, if $\boldsymbol{b}_i \cap \boldsymbol{b}_j = \boldsymbol{b}_i$ and $area(\boldsymbol{b}_i) < \gamma_2 * area(\boldsymbol{b}_j)$, where $area(\boldsymbol{b_k})$ is defined as the area of a box $\boldsymbol{b}_k$ and $\gamma_2 < 0.5$. Otherwise, it returns $False$. In other words, when $\boldsymbol{b}_i$ is greatly included by $\boldsymbol{b}_j$, they have 'deep inclusion' relation.

Given K instance refinement branches after MIL branch, proposals in K-th refinement often contain more global regions than those in the first refinement, which is the effect of multiple refinement branches. Based on this effect, we select the top1-scored proposal of target class c in the first and K-th refinement as $\boldsymbol{p}_{refine_1}$ and $\boldsymbol{p}_{refine_K}$. We then compare the corresponding boxes of the two proposals, if $deep\_inclusion(\boldsymbol{b}_{refine_1}, \boldsymbol{b}_{refine_K}) = True$, then $\boldsymbol{p}_{refine_1}$ is regarded as salient region proposal of class c, as illustrated in Figure 6(a). Following this manner, salient region proposals of all classes in Y can be collected.

We then create a Salient Region Bank $Bank[c] = \{\{\boldsymbol{np}_1^c, ..., \boldsymbol{np}_R^c\}, \{ns_1^c, ..., ns_R^c\}\}$ for each class c to preserve and update salient region proposals $\boldsymbol{np}_i^c$ with their predicted score ns_i^c. The updating rule follows Wang *et al.* [5]. Figure 6(b) show cases selected from the Salient Region Banks, both from early training iterations and the final training iteration. It is evident that throughout the training process, the proposals consistently chosen are salient region proposals.

Contrastive learning can be implemented to push training proposals away from proposals in salient region banks. Following Wang *et al.* [5], we minimize the maximum similarity between a proposal $\boldsymbol{p}_j$ and salient region proposals of its predicted class c_j using Equation 5.

$$\mathcal{L}_{contrast} = \Sigma_{j=1}^{|P|} ns_{i*}^{c_j} max_i \left(\frac{\boldsymbol{p}_j \cdot \boldsymbol{np}_i^{c_j}}{\|\boldsymbol{p}_j\| \cdot \|\boldsymbol{np}_i^{c_j}\|}\right) \quad (5)$$

III. Experiments

A. Datasets, Metrics and Implementaion Details

We evaluate our proposed intra-proposal and inter-proposal approaches on PASCAL VOC 2007, VOC 2012 [13] and MS COCO [14] benchmarks. For VOC datasets, trainval and test sets are employed for training and testing respectively, and two metrics are used to evaluate the performance: mean average precision (mAP) to evaluate model performance on test set, Correct Localization (CorLoc) to evaluate localization capability on trainval set. For COCO datasets, we adopt the train2014 and val2014 sets for training and testing, and use AP and AP50 for performance evaluation.

Following previous WSOD methods, we utilize VGG16 model pre-trained on ImageNet dataset [15] as our backbone network and utilize a regression branch after the last OIR branch. A 12-layer Vision Transformer encoder with self-attention modules is adopted to implement PFL-WSOD_$intra$, and we load parameters pretrained on ImageNet dataset [15]. MCG [16] is adopted as the external proposal generation method. During training, the mini-batch size is set to 4, and the learing rate for VOC 2007, VOC 2012 and COCO are set to 0.0005 for the first 25K, 60k, 100K iterations and decreased by a factor of 0.1 for the following 10K, 10K, 80k iterations respectively. The momentum is set to 0.9. High score proposals are kept and Non-Maximum Suppression (with 30 % IOU threshold) is applied to calculte mAP and CorLoc. For PFL-WSOD_$intra$, the loss function is $\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{OIR}$. For PFL-WSOD_$inter$, the loss function is $\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{OIR} + \mathcal{L}_{contrast}$. Our experiments are conducted using an NVIDIA GeForce RTX 3090 GPU.

B. Quantitative and qualitative results

In this section, we compare our proposed approaches with state-of-the-arts on WSOD benchmarks. For both proposed methods, we utilize Negative Guided Instance Selection (NGIS) module [5] to further relieve the *missing instance* problem by replacing their incorrect instances with our Salient Region Banks described in Section 2.2.2.

From Table I we can see that our PFL-WSOD_$intra$ achieves 57.2% on mAP and 72.2% on CorLoc, and our PFL-WSOD_$inter$ achieves 57.0% on mAP and 72.0% on CorLoc, which both outperform existing baselines. Furthermore, by replacing the proposal generation method from MCG to COB [17], our approaches obtain further performance improvements and achieve state-of-the-art as shown by the PFL-WSOD_$intra^*$ and PFL-WSOD_$inter^*$. Specifically, we obtain 58.3% on mAP for both models.

To further demonstrate the superiority of our methods, we evaluate our method on VOC 2012 as shown in Table II. Our PFL-WSOD_$inter$ achieves remarkable performances on VOC 2012, with 5.8% improvement on mAP and 5.3% improvement on CorLoc compared with the most State-of-the-Art method. The performance is further improved when using a different proposal generation method [17], which is 60.4% on mAP and 78.5% on CorLoc.

The results of our two methods on MS COCO is illusteated in Table III. PFL-WSOD_$intra$ produces high performances of 24.8% on AP50 and 11.6% on AP, and PFL-WSOD_$inter$ also presents satisfying performances of 25.5% on AP50 and 12.0% on AP, both outperforming ICM-WSOD [4], which also validates the effectiveness of our PFL-WSOD.

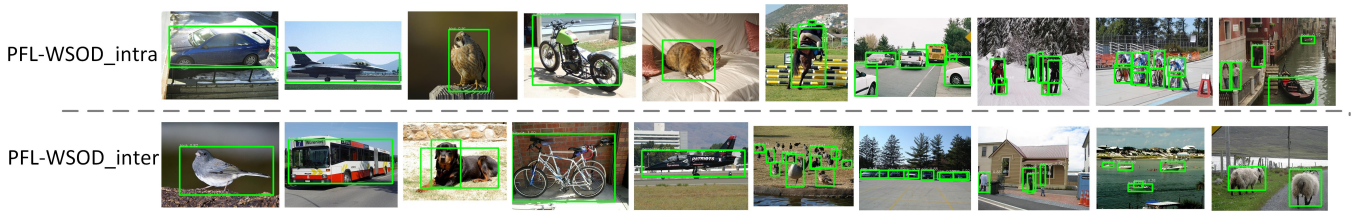

Fig. 7. Visualized results of PFL-WSOD_$intra$ and PFL-WSOD_$inter$. Boxes with confidence larger than 0.3 are selected as the prediction.

Method	mAP (%)	CorLoc (%)
OICR [3]	41.2	60.6
PCL [18]	43.5	62.7
C-MIDN [19]	52.6	68.7
WSOD2 [20]	53.6	69.5
OIM [21]	50.1	67.2
SLV [22]	53.5	49.2
ICM-WSD [4]	54.9	68.8
CASD [23]	56.8	70.4
NDI-WSOD [5]	56.8	71.0
PFL-WSOD_$intra$ (ours)	57.2	72.2
PFL-WSOD_$intra^*$ (ours)	**58.3**	72.8
PFL-WSOD_$inter$ (ours)	57.0	72.0
PFL-WSOD_$inter^*$ (ours)	**58.3**	**73.0**

TABLE I
COMPARISON WITH THE STATE-OF-THE-ARTS IN MAP AND CORLOC ON VOC2007 *test* AND *trainval* RESPECTIVELY.

Method	mAP (%)	CorLoc (%)
OICR [3]	37.9	62.1
PCL [18]	40.6	63.2
C-MIDN [19]	50.2	71.2
WSOD2 [20]	47.2	71.9
OIM [21]	45.3	67.1
SLV [22]	53.5	49.2
ICM-WSD [4]	52.1	70.9
CASD [23]	53.6	72.3
NDI-WSOD [5]	53.9	72.2
PFL-WSOD_$intra$ (ours)	53.6	73.2
PFL-WSOD_$intra^*$ (ours)	54.7	74.1
PFL-WSOD_$inter$ (ours)	59.7	77.5
PFL-WSOD_$inter^*$ (ours)	**60.4**	**78.5**

TABLE II
COMPARISON WITH THE STATE-OF-THE-ARTS IN MAP AND CORLOC ON VOC2012 *test* AND *trainval* RESPECTIVELY.

Figure 7 provides visualizations of our detection on VOC 2007 test set. We can see that both methods produce satisfactory bounding boxes on rigid objects (bus, aeroplane, bicycle and boat) and animals (bird, dog, cat and horse), therefore relieving the *salient region problem*. Also, both methods successfully detect multiple objects, thus relieving the *missing instance* problem.

C. Analysis on Similarity($salient, global$)

We first introduce the details of how to calculate Similarity($salient, global$). Given a image, we first annotate a salient region proposal $\boldsymbol{p}_{annotate}$ manually. Then, a group of salient region proposals are collected by selecting proposals having high overlaps with $\boldsymbol{p}_{annotate}$. A salient feature representation vector $\boldsymbol{p}_{salient}$ is obtained by averaging these collected proposals. Following the same procedures, we obtain a global feature representation vector $\boldsymbol{p}_{global}$. Finally, cosine similarity is calculated between $\boldsymbol{p}_{salient}$ and $\boldsymbol{p}_{global}$.

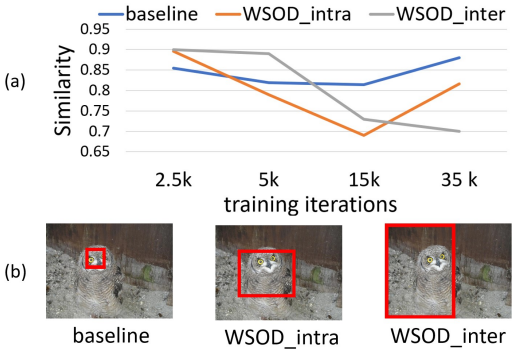

Fig. 8. (a) Comparison of Similarity($salient, global$) of the training image in Figure 3 using baseline and our methods. (b) Top1 score proposal detected by baseline and our methods in the final training iteration.

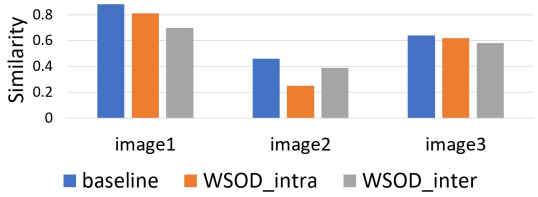

Fig. 9. Comparison of Similarity($salient, global$) of different training image using baseline and our methods.

We then analyze Similarity($salient, global$). We calculate Similarity($salient, global$) of the training image of Figure 3 in different training iterations for baseline and our methods. As shown in Figure 8(a), our methods obtain lower Similarity($salient, global$) than the baseline, especially after 15K iterations. Correspondingly, the top1-scored proposal detected by MIL branch is also better than the baseline, as shown in Figure 8(b). We further compare Similarity($salient, global$) of random training images at the final iteration. As shown in Figure 9, Similarities($salient, global$) on our methods are consistently lower than baseline. Therefore, PFL-WSOD effectively reduces Similarity($salient, global$), enabling the detector to effectively differentiate between salient region proposals and global region proposals. This ultimately leads

Method	AP	AP50	Method	AP	AP50
PCL	8.5	19.4	ICM-WSD	11.4	24.3
C-MIDN	9.6	21.4	PFL-WSOD_$intra$	11.6	24.8
WSOD2	10.8	22.7	PFL-WSOD_$inter$	**12.0**	**25.5**

TABLE III
COMPARISON ON THE MS COCO DATASET.

Method	mAP(%)
Baseline	52.9
PFL-WSOD_$intra$ (w/o NGIS)	56.0
PFL-WSOD_$inter$ (w/o NGIS)	56.0
PFL-WSOD_$intra$	**57.2**
PFL-WSOD_$inter$	57.0

TABLE IV
ABLATION STUDY ON VOC 2007 TEST SET.

to enhanced detection performance under weak supervision.

D. Ablation Study

We perform ablation studies for our proposed methods on VOC 2007 test set. We can see from Table IV that both PFL-WSOD_$intra$ and PFL-WSOD_$inter$ achieve 56% on mAP without utilizing NGIS module , where each achieves 3.1% improvement against the baseline. These results illustrate that our proposed Self-Attention on Single Proposal and contrastive learning with Salient Region Banks learn better proposal features. Additionally, when integrated with NGIS, both of our proposed methods exhibit enhanced performance gains with 1.2% improvement for PFL-WSOD_$intra$ and 1.0% improvement for PFL-WSOD_$inter$ over the baseline respectively.

IV. CONCLUSION

In this paper, we propose a novel WSOD framework named PFL-WSOD for learning better proposal features. We utilize two types of proposal relations to implement PFL-WSOD, $i.e.$, PFL-WSOD_$intra$ and PFL-WSOD_$inter$. While the former method utilizes self-attention on single proposal, the latter method combine contrastive learning with salient region banks. Our methods achieve state-of-the-art performances.

REFERENCES

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

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

[3] P. Tang, X. Wang, X. Bai, and W. 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.

[4] Z. Ren, Z. Yu, X. Yang, M.-Y. Liu, Y. J. Lee, A. G. Schwing, and J. Kautz, "Instance-aware, context-focused, and memory-efficient weakly supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 598–10 607.

[5] G. Wang, X. Zhang, Z. Peng, X. Tang, H. Zhou, and l. Jiao, "Absolute wrong makes better: Boosting weakly supervised object detection via negative deterministic informations," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022, pp. 1378–1384.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] M. Liao, F. Wan, Y. Yao, Z. Han, J. Zou, Y. Wang, B. Feng, P. Yuan, and Q. Ye, "End-to-end weakly supervised object detection with sparse proposal evolution," in *European Conference on Computer Vision*. Springer, 2022, pp. 210–226.

[8] T. LaBonte, Y. Song, X. Wang, V. Vineet, and N. Joshi, "Scaling novel object detection with weakly supervised detection transformers," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 85–96.

[9] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 213–229.

[10] W. Zhang, X. Zhang, M.-L. Zhang *et al.*, "Multi-instance causal representation learning for instance label prediction and out-of-distribution generalization," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34 940–34 953, 2022.

[11] W. Tang, W. Zhang, and M.-L. Zhang, "Disambiguated attention embedding for multi-instance partial-label learning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[12] Y. Xu, L. Zhang, and D. Zhou, "Teca: A two-stage approach with controllable attention soft prompt for few-shot nested named entity recognition," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 15 698–15 710.

[13] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–308, 2009.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.

[16] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 328–335.

[17] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool, "Convolutional oriented boundaries," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 580–596.

[18] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. 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.

[19] Y. Gao, B. Liu, N. Guo, X. Ye, F. Wan, H. You, and D. Fan, "C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9834–9843.

[20] Z. Zeng, B. Liu, J. Fu, H. Chao, and L. Zhang, "Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8292–8300.

[21] C. Lin, S. Wang, D. Xu, Y. Lu, and W. Zhang, "Object instance mining for weakly supervised object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 482–11 489.

[22] Z. Chen, Z. Fu, R. Jiang, Y. Chen, and X.-S. Hua, "Slv: Spatial likelihood voting for weakly supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 995–13 004.

[23] Z. Huang, Y. Zou, B. Kumar, and D. Huang, "Comprehensive attention self-distillation for weakly-supervised object detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 797–16 807, 2020.