WEAKLY- AND SEMI-SUPERVISED OBJECT LOCALIZATION

Zhen-Tang Huang, Yan-He Chen, Mei-Chen Yeh

National Taiwan Normal University

ABSTRACT

Weakly supervised object localization deals with the lack of location-level labels to train localization models. Recently a new evaluation protocol is proposed in which full supervision is available but limited to only a small validation set. It motives us to explore semi-supervised learning for addressing this problem. In particular, the localization model is developed via self-training: we use a small amount of data with full supervision to train a class-agnostic detector, and use it to generate pseudo bounding boxes for data with weak supervision. Furthermore, we propose a selection algorithm to discover high-quality pseudo labels, and deal with data imbalance caused by pseudo labeling. We demonstrate the superiority of the proposed method with performance on par with the state of the art on two benchmarks.

Index Terms— weakly supervised object localization, semi-supervised learning, deep learning

1. INTRODUCTION

With the success of deep neural networks in object detection, weakly-supervised object localization (WSOL) has recently received much attention because it alleviates a huge amount of human efforts to annotate training samples. Different to fully-supervised object localization that requires location-level labels (e.g., bounding boxes) of each training instance, WSOL has only image-level labels. A WSOL model is trained to localize objects of interest under the setting in which only class labels are given. WSOL significantly reduces the data annotation cost, being essential to numerous real-world applications where fully-supervised data are difficult to collect, including autonomous driving [1] and defect detection in industrial inspection [2]—just name a few.

Recently, Choe *et al.* [3] find the WSOL problem ill-posed with only image-level labels and propose an evaluation protocol where a small validation set with full supervision is available. This study also shows that recent WSOL methods have not made a major improvement over the class activation mapping (CAM) baseline [3]. More interestingly, a few-shot learning method outperforms existing WSOL methods, where the full-supervision at validation time is used for model training instead.

Considering that a small amount of data with full supervision and a large amount of data with weak supervision are both available for training a WSOL model, we propose a new approach that explores semi-supervised learning (SSL) to tackle this problem. We first train an object detection model using the validation set, in which the bounding boxes of samples are given, and then use the training set to perform self-training. One benefit of this SSL method is that modern fully-supervised object detection techniques can be applied to WSOL. However, learning a robust WSOL model with a SSL setting introduces new challenges as well. First, the object detection model—trained with *few data*—must reach a reasonable performance to infer pseudo labels from weakly labeled data. Second, the model will be trained inevitably with *imbalanced data* because the number of pseudo labels inferred by the model may differ significantly for each class. In particular, the amount of pseudo labels created from difficult samples would be small if we screen the pseudo labels by quality (e.g., confidence score of the detection).

In this paper, we address the above issues and present a SSL method for WSOL. This method provides an effective alternative to the few-shot learning baseline. With promising object localization performance under semi-supervised scenarios, we further investigate the generalization capability of the proposed method. The proposed method—trained with the source domain data (e.g., ImageNet [4])—achieves the state-of-the-art performance when it is directly evaluated using the target domain data (e.g., CUB [5]) *without fine-tuning*.

We summarize the main contributions as follows: (1) We propose a SSL based method to address the WSOL problem. To the best of our knowledge, we are among the first that explore SSL to tackle this problem. (2) We identify the challenges in developing a SSL based method for WSOL, including the training of a robust base detector using a few labeled samples per class and training with long-tailed distributed data caused by pseudo labeling. The proposed method addresses both issues. (3) We evaluate the proposed method on WSOL benchmarks and show that it improves previous methods by a large margin in localization accuracy. We further conduct a cross-dataset evaluation to demonstrate its generalization capability.

This work was supported by the National Science and Technology Council of Taiwan (110-2221-E-003-016, 110-2634-F-002-050).

2. RELATED WORK

In this section, we describe related works to this study, including weakly supervised object localization and semi-supervised object detection.

Since the seminal WSOL work of class activation mapping (CAM) [6], the field has focused on how to expand the attention regions to cover objects more broadly and localize them better. For example, adversarial complementary learning (ACoL) [7] and multiple erasing integrated learning (MEIL) [8] erase the most discriminative part in each training iteration, forcing the network to pay attention to different regions besides the most discriminative one. Zhang *et al.* refine the CAM map using Weighted Global Average Pooling and adaptively select a threshold to achieve better object localization [9]. Pseudo supervised object localization (PSOL) [10] uses deep descriptor transformation to generate noisy pseudo annotations and then perform bounding box regression on them. Recently, shallow feature-aware pseudo supervised object localization (SPOL) [11] generates the CAMs through a novel element-wise multiplication of shallow and deep feature maps, which filters the background noise and generates sharper boundaries. Unlike these mainstream methods, we address the WSOL problem by semi-supervised learning.

A majority of the semi-supervised object detection models are developed based on Faster R-CNN [12]. For example, STAC [13] uses labeled data to train the detection model that subsequently infers pseudo labels from unlabeled data. The model is then trained with both labeled and unlabeled data. Recent methods apply an end-to-end framework [14, 15, 16], in which a teacher model and a student model are jointly trained in a mutually-beneficial manner. Although these methods achieve state-of-the-art detection performance on object detection benchmarks, they are not computationally feasible to the WSOL problem in which a large-scale weakly labeled image set is available for training. The process of labeling a large dataset is costly for methods that involve iterative re-training. Therefore, we apply a simple self-training pipeline and will show that it is a strong competitor to state-of-the-art methods in Sec. 4.

3. METHOD

The overview of our method is displayed in Fig. 1. We decouple the WSOL task into class-agnostic object detection and image classification, alleviating the requirement of abundant samples per class to train a robust detector. First, we train a class-agnostic detection model with the validation set containing only a few location-level labeled samples (Fig. 1(a)). This step is identical to any supervised learning method. Next, we perform self-training via pseudo labeling—the detection model is used to generate the pseudo bounding boxes for all training images, which contain only image-level annotations. The process is shown in Fig. 1(b). In particular, we propose a

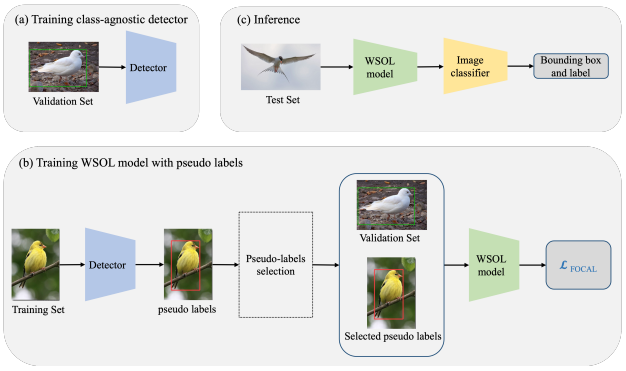

Fig. 1. Model architecture. Please see texts for details.

selection scheme to find reliable pseudo labels. The detection model is re-trained by using both of the fully and pseudo labeled images. Finally, we train an image classifier using the training set to determine the class label of an image. During inference, a test image is processed by the detector to obtain bounding boxes and then is classified by the classifier to obtain the label of each bounding box (Fig. 1(c)).

The quality is as important as the quantity of pseudo labels. Using low-quality pseudo labels might hinder the model from effective detection. However, selecting pseudo labels may lead to data imbalance: some categories may contain more data samples than the others. We will address these issues in the following subsections.

3.1. Class-Agnostic Detection

The base detector model must reach a good detection accuracy because it will be used to infer the pseudo labels from the training set. However, only a small amount of labeled data is available to train the base detector. For example, the ImageNet validation dataset contains 10 labeled samples per class, and the CUB validation dataset contains 5 labeled samples per class [3]. It is difficult to train a general object detection model with such a small amount of labeled data.

We develop the base detection model upon YOLOv5 [17], which is a modern model for fully supervised object detection. We train a *class-agnostic* detector, in which the model aims to localize foreground objects. With the labeled data in the validation set, it is difficult to train YOLO to simultaneously perform localization and classification. Moreover, the localization task requires the extraction of global features from the whole object, while the classification task often relies on the most discriminate part (local features) of the object. Therefore, training a foreground object detector can not only alleviate the problem of insufficient location-level labeled data but also reduce the task complexity by excluding the classification objective. We focus on the capability of the model to localize foreground objects, subsequently applied to produce pseudo labels from weakly labeled training data.

Algorithm 1 Selecting Reliable Pseudo Labels

Input:

Samples $\mathcal{S}$: each sample x_i has a class label y_i, n_i pseudo boxes $\{\mathbf{b}_j\}_{j=1}^{n_i}$ and n_i confidence scores $\{s_j\}_{j=1}^{n_i}$;
A threshold γ;
The number of selected samples per class N;

Output:

Selected samples $\mathcal{F}$;

1: Initialize $\mathcal{F} \leftarrow \emptyset$
2: **for** $x_i \in \mathcal{S}$ **do**
3: Calculate $m_i \leftarrow \min_j\{s_j\}$;
4: **end for**
5: Sort samples by descending order of m_i
6: **for** $x_i \in$ sorted $\mathcal{S}$ **do**
7: let n be the no. of selected samples whose label is y_i;
8: **if** $m_i \geq \gamma$ and $n < N$ **then**
9: Put x_i into $\mathcal{F}$;
10: **end if**
11: **end for**
12: **return** $\mathcal{F}$;

3.2. Pseudo-Labels Selection

The usage of pseudo labels is important in semi-supervised learning [18]. Specifically, the trade-off between the quantity and the quality of pseudo labels must be considered. The proposed pseudo-labels selection algorithm prioritizes high-confidence pseudo labels, avoiding the involvement of poor-quality ones that may deteriorate the model training and affect accuracy. The pseudo-labels selection algorithm is summarized in Algorithm 1.

Algorithm 1 receives the training image set $\mathcal{S}$, a threshold value γ and the number of required samples per class N. Each training sample has n_i pseudo bounding boxes $\{\mathbf{b}_j, s_j\}_{j=1}^{n_i}$—each bounding box $\mathbf{b}_j$ has a confidence score s_j. The threshold γ determines the quality requirement for the pseudo bounding boxes. The output set $\mathcal{F}$ contains the selected training images and their pseudo bounding boxes.

The selection is performed based on the confidence score of a detected box returned by the detector. If the confidence scores of *all* detected boxes are greater than the threshold, the sample is selected. Note that the step of calculating the minimal confidence score for all bounding boxes in one image (line 3 in Algorithm 1) is important, which is not as trivial as that in a classification task. In object localization we may have more than one instance in one sample. If we simply select the instances with large confidence scores, we may feed the model *with partial annotation*. This is hurtful to model training because some foreground objects are treated to be background. Furthermore, as we have a large set of weakly labeled training samples, we can take a strict selection policy—only those images whose bounding boxes scores are all greater than the threshold are selected.

3.3. Model Training and Inference

Despite that pseudo labeling is simple yet effective for utilizing unlabeled samples, it creates the data imbalance problem because classes have different amount of hard samples. For some classes that are more difficult to detect, the number of selected samples is small, which exacerbates the difficulty of training the model for detecting those classes. Therefore, the WSOL model is trained with the focal loss [19], which can focus more on difficult samples during training.

The inference procedure is displayed in Fig. 1(c). We feed a test image into the WSOL model to obtain bounding boxes, and then use an image classifier to compute the class label for each box. Specifically, we use ResNeXt101 [20] and NTSNET [21] models to train the image classifiers for ImageNet [4] and CUB [5], respectively.

4. EXPERIMENTS

4.1. Experimental Setting

We used two benchmark datasets in the experiments, including ImageNet [4] and Caltech-UCSD Birds-200-2011 (CUB) [5]. We followed the protocol specified in [3] for dividing the data for training, validation and evaluation.

Three evaluation metrics—GT-known localization accuracy (GT Loc), top-1 localization accuracy (Top-1 Loc) and top-5 localization accuracy (Top-5 Loc)—were applied to evaluate the performance of the proposed method. This study uses GT Loc as the main evaluation metric because the essence of WSOL is localization rather than classification. This echos the claim made in [3], in which the authors advocate the measurement of localization performance alone.

We trained our model for 300 epochs with a batch size 30. In the ImageNet experiment N was set to 20, resulting in a total of 19,032 pseudo-labeled images to train the WSOL model. The confidence threshold was set to 0.9. In the CUB experiments N was set to 10, resulting in a total number of 1,518 images to train the WSOL model. The confidence threshold was set to 0.95. All experiments were conducted on a machine with an NVIDIA GeForce RTX 3090 GPU.

4.2. Experimental Results

We compared the proposed method with recent state-of-the-art WSOL methods [10, 22, 23, 11, 26, 9, 3, 24, 25]. The results on ImageNet and CUB are shown in Table 1 and Table 2, respectively. We reported the performance of the supervised baseline, which used only the validation set for training the model, and two variants of the proposed SSL based method. The first variant used the cross entropy loss to optimize the model, and the second variant used the focal loss.

We first examine the effectiveness of using weakly-labeled data. The proposed SSL method improves the supervised baseline no matter which loss function is applied.

Table 1. Comparison with state-of-the-arts on ImageNet.

Method	GT Loc	Top-1 Loc	Top-5 Loc
FSL [3] (*CVPR*'20)	66.30	-	-
PSOL [10] (*CVPR*'20)	66.28	55.31	64.18
SLT-Net [22] (*CVPR*'21)	67.60	55.70	65.40
SCG [23] (*CVPR*'21)	65.05	49.56	61.32
SPOL [11] (*CVPR*'21)	69.02	**59.14**	**67.15**
Zhang *et al.* [9] (*ICASSP*'22)	65.40	50.10	-
Kim *et al.* [24] (*CVPR*'22)	69.89	53.76	65.75
Zhu *et al.* [25] (*CVPR*'22)	70.27	55.84	-
Wu *et al.* [26] (*CVPR*'22)	72.00	52.97	66.59
Supervised baseline	61.45	46.73	56.44
SSL w. cross entropy loss	67.57	50.35	61.22
SSL w. focal loss	**74.72**	54.09	66.51

Table 2. Comparison with state-of-the-arts on CUB.

Method	GT Loc	Top-1 Loc	Top-5 Loc
FSL [3] (*CVPR*'20)	92.00	-	-
PSOL [10] (*CVPR*'20)	93.01	77.44	89.51
SLT-Net [22] (*CVPR*'21)	87.60	67.80	-
SCG [23] (*CVPR*'21)	72.14	53.59	66.50
SPOL [11] (*CVPR*'21)	96.46	80.12	**93.44**
Zhang *et al.* [9] (*ICASSP*'22)	82.32	61.85	-
Kim *et al.* [24] (*CVPR*'22)	93.17	70.83	88.07
Zhu *et al.* [25] (*CVPR*'22)	81.83	66.65	-
Wu *et al.* [26] (*CVPR*'22)	95.13	77.25	90.08
Supervised baseline	96.81	79.70	91.77
SSL w. cross entropy loss	97.96	**80.89**	92.80
SSL w. focal loss	**98.39**	79.96	92.77
Supervised baseline*	95.75	79.55	91.04
SSL w. cross entropy loss*	95.01	78.72	90.23
SSL w. focal loss*	96.05	79.57	91.15

For example, the model trained with the focal loss achieves 74.72% GT Loc while the baseline achieves 61.45% in the ImageNet experiment. The same observation on performance improvement can also be made in the CUB experiment (98.39% vs. 96.81%). Next, using focal loss can improve the localization performance. For example, the model improves 7.15% in GT Loc when it is trained with the focal loss in the ImageNet experiment. In CUB, the performance gain is also observed but is not significant. In comparison with the state-of-the-art methods, the proposed method trained with the focal loss achieves the best performance in GT Loc in both experiments. The result validates the effectiveness of the SSL strategy that trains the model using the validation data and re-trains it using the training data with pseudo labels. Using the focal loss can effectively address data imbalance and sample difficulty, which is a challenge we must address when applying the pseudo labeling strategy for WSOL. Although we do not focus on classification in the proposed method, our performances on top-1 Loc and top-5 Loc are competitive to those of state-of-the-arts. Moreover, our model is trained by using only a small amount of training samples (those selected by the proposed algorithm). For example, we only use 2.3% training data in ImageNet and 42% training data in CUB. Table 2 shows the performance of our method *trained on ImageNet*, denoted by adding a asterisk symbol (*). Without using any data on CUB, our method trained with the focal loss achieves 96.05% GT Loc and outperforms many state-of-the-art methods. A few localization results are displayed in Fig. 2. The proposed method has a good localization performance with different number of instances, illumination conditions and complex background.

Table 3. Analysis of reliable bounding box selection

Method	min. conf.	GT Loc	Top-1	Top-5
SSL w. cross entropy	✗	64.95	48.81	59.12
	✓	67.57	50.35	61.22
SSL w. focal loss	✗	72.76	53.22	65.11
	✓	74.72	54.09	66.51

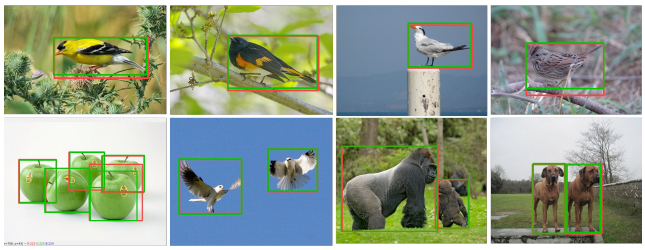

Fig. 2. A few localization results of the proposed method. Top: CUB, Bottom: ImageNet; Green: GT, Red: Ours.

Finally we analyze the effect of manners for selecting reliable bounding boxes. Recall that in Algorithm 1 we computed the minimum score of all detected bounding boxes in one image and required that a selected sample should have all bounding boxes with scores greater than the threshold. Now we perform an alternative: we simply threshold on each bounding box, without requiring that all bounding boxes in one image must be used. The results are shown in Table 3. The accuracy of the model applying the proposed selection strategy is higher than that of the alternative. The results shows that it is important to select pseudo labels carefully, as a trivial solution may lead to a situation where some instances may be regarded as background and thereby achieve an inferior performance.

5. CONCLUSION

We present a new WSOL method based on SSL. In this approach, we utilize the validation set to train a base model, and then use it to explore the training set via pseudo labeling. Extensive experimental results validate the design choices of the proposed method. One possible future direction is to elaborate pseudo labels more effectively; for example, we may consider the confidence score and the stability.

6. REFERENCES

[1] Aseem Behl *et al.*, "Bounding boxes, segmentations and object coordinates: How important is recognition for 3d scene flow estimation in autonomous driving scenarios?," in *IEEE Conference on Computer Vision*, 2017.

[2] Liang Xu *et al.*, "A weakly supervised surface defect detection based on convolutional neural network," *IEEE Access*, vol. 8, pp. 42285–42296, 2020.

[3] Junsuk Choe *et al.*, "Evaluation for weakly supervised object localization: Protocol, metrics, and datasets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[4] Olga Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[5] Peter Welinder *et al.*, "Caltech-ucsd birds 200," 2010.

[6] Bolei Zhou *et al.*, "Learning deep features for discriminative localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[7] Xiaolin Zhang *et al.*, "Adversarial complementary learning for weakly supervised object localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[8] Jinjie Mai, Meng Yang, and Wenfeng Luo, "Erasing integrated learning: A simple yet effective approach for weakly supervised object localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[9] Zhenfei Zhang, Ming-Ching Chang, and Tien D. Bui, "Improving class activation map for weakly supervised object localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[10] Chen-Lin Zhang, Yun-Hao Cao, and Jianxin Wu, "Rethinking the route towards weakly supervised object localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[11] Jun Wei *et al.*, "Shallow feature matters for weakly supervised object localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[12] Shaoqing Ren *et al.*, "Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015.

[13] Kihyuk Sohn *et al.*, "A simple semi-supervised learning framework for object detection," *arXiv preprint arXiv:2005.04757*, 2020.

[14] Yihe Tang *et al.*, "Humble teachers teach better students for semi-supervised object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[15] Qiang Zhou *et al.*, "Instant-teaching: An end-to-end semi-supervised object detection framework," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[16] Yen-Cheng Liu *et al.*, "Unbiased teacher for semi-supervised object detection," in *International Conference on Learning Representations*, 2021.

[17] Glenn Jocher, "ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements," https://github.com/ultralytics/yolov5, Oct. 2020.

[18] Lihe Yang *et al.*, "St++: Make self-training work better for semi-supervised semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[19] Tsung-Yi Lin *et al.*, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision*, 2017.

[20] Saining Xie *et al.*, "Aggregated residual transformations for deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[21] Ze Yang *et al.*, "Learning to navigate for fine-grained classification," in *European Conference on Computer Vision*, 2018.

[22] Guangyu Guo *et al.*, "Strengthen learning tolerance for weakly supervised object localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[23] Xingjia Pan *et al.*, "Unveiling the potential of structure preserving for weakly supervised object localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[24] Eunji Kim *et al.*, "Bridging the gap between classification and localization for weakly supervised object localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[25] Lei Zhu *et al.*, "Weakly supervised object localization as domain adaption," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[26] Pingyu Wu, Wei Zhai, and Yang Cao, "Background activation suppression for weakly supervised object localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.