End-to-End Weakly Supervised Object Detection with Sparse Proposal Evolution

Mingxiang Liao[1], Fang Wan[1] ⋆, Yuan Yao[1], Zhenjun Han[1], Jialing Zou[1], Yuze Wang[2], Bailan Feng[2], Peng Yuan[2], and Qixiang Ye[1]

[1] University of Chinese Academy of Sciences, Beijing, China
[2] Huawei Noah's Ark Lab
{liaomingxiang20,yaoyuan17}@mails.ucas.ac.cn,
{wanfang,hanzhj,qxye}@ucas.ac.cn, zjl7223009@163.com,
{wangyuze1,fengbailan,yuanpeng126}@huawei.com

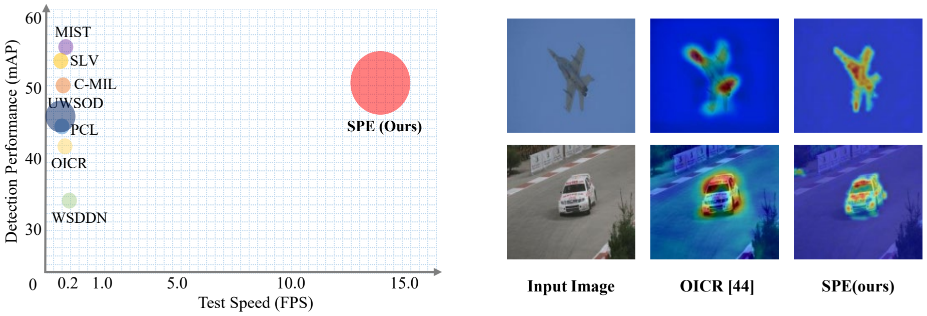
Abstract. Conventional methods for weakly supervised object detection (WSOD) typically enumerate dense proposals and select the discriminative proposals as objects. However, these two-stage "enumerate-and-select" methods suffer object feature ambiguity brought by dense proposals and low detection efficiency caused by the proposal enumeration procedure. In this study, we propose a sparse proposal evolution (SPE) approach, which advances WSOD from the two-stage pipeline with dense proposals to an end-to-end framework with sparse proposals. SPE is built upon a visual transformer equipped with a seed proposal generation (SPG) branch and a sparse proposal refinement (SPR) branch. SPG generates high-quality seed proposals by taking advantage of the cascaded self-attention mechanism of the visual transformer, and SPR trains the detector to predict sparse proposals which are supervised by the seed proposals in a one-to-one matching fashion. SPG and SPR are iteratively performed so that seed proposals update to accurate supervision signals and sparse proposals evolve to precise object regions. Experiments on VOC and COCO object detection datasets show that SPE outperforms the state-of-the-art end-to-end methods by 7.0% mAP and 8.1% AP50. It is an order of magnitude faster than the two-stage methods, setting the first solid baseline for end-to-end WSOD with sparse proposals. The code is available at github.com/MingXiangL/SPE.

Keywords: Weakly Supervised Object Detection, Sparse Proposals, Proposal Evolution, End-to-end Training

1 Introduction

Visual object detection has achieved unprecedented progress in the past decade. However, such progress heavily relies on the large amount of data annotations (*e.g.*, object bounding boxes) which require extensive human effort and time cost. Weakly supervised object detection (WSOD), which only requires image-level annotations indicating the presence or absence of a class of objects, significantly reduces the annotation cost [4,31,53,14,32,33].

[⋆] Corresponding author.



(a) Comparison of detection efficiency.

(b) Comparison of activation maps.

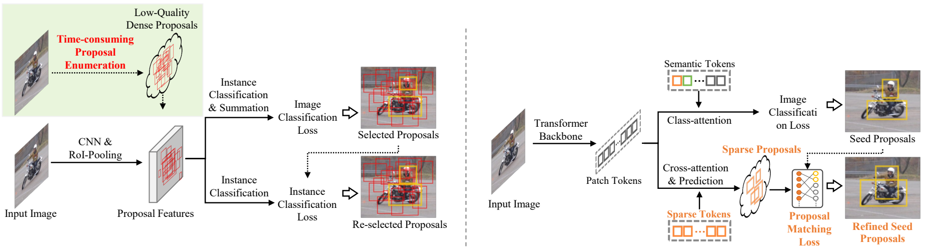
Fig. 1: Comparison of (a) detection efficiency and (b) activation maps between the conventional methods and the proposed SPE for weakly supervised object detection (WSOD) on VOC 2007. In (a), larger cycles denote higher proposal generation speeds. All speeds in (a) are evaluated on a NVIDIA RTX GPU.

For the lack of instance-level annotation, WSOD methods require to localize the objects while estimate object detectors at the same time during training. To fulfill this purpose, the early WSDDN method [6] used an "enumerate-and-select" pipeline. It firstly enumerates dense proposals using empirical clues [37,27] to ensure a high recall rate and then selects the most discriminative proposal as the pseudo object for detector training. Recent studies improved either the proposal enumeration [45,42,38] or the proposal selection module [44,19,52,23].

However, this "enumerate-and-select" pipeline meets the performance upper bound for the following two problems: (1) The redundant and near-duplicate proposals aggregate the difficulty to localize objects and decrease the detection efficiency, Fig. 1(a). (2) During training, the labels of the dense proposals are assigned by a single pseudo object through a many-to-one matching strategy, *i.e.*, multiple proposals with large IoUs between the pseudo object are selected for detector training, which introduces ambiguity to feature representation, Fig. 1(b).

In this paper, we propose the sparse proposal evolution (SPE) approach, which advances WSOD from the enumerate-and-select pipeline with dense proposals (Fig. 2(a)) to an end-to-end framework with sparse proposals (Fig. 2(b)). SPE adopts a "seed-and-refine" approach, which first produces sparse seed proposals and then refines them to achieve accurate object localization.

SPE consists of a seed proposal generation (SPG) branch and a sparse proposal refinement (SPR) branch. During training, SPG leverages the visual transformer [47] to generate semantic-aware attention maps. By taking advantage of the cascaded self-attention mechanism born with the visual transformer, the semantic-aware attention map can extract long-range feature dependencies and activate full object extentFig. 1(b). With these semantic-aware attention maps, SPG can generate high-quality seed proposals. Using the seed proposals as pseudo supervisions, SPR trains a detector by introducing a set of sparse



(a) **Enumerate-and-Select Pipeline:** WSDDN [6], OICR [44]

(b) **Seed-and-Refine Framework**: SPE (Ours)

Fig. 2: Comparison of (a) the conventional "enumerate-and-select" pipeline with (b) our "seed-and-refine" framework for weakly supervised object detection.

proposals that are learned to match with the seed proposals in a one-to-one matching fashion. During the proposal matching procedure, each seed proposal is augmented to multiple orientations, which provide the opportunity to refine object locations when the proposals and the detector evolve.

The contributions of this study include:

- We propose the sparse proposal evolution (SPE) approach, opening the promising direction for end-to-end WSOD with sparse proposals.
- We update many-to-one proposal selection to one-to-one proposal-proposal matching, making it possible to apply the "seed-and-refine" mechanism in the challenging WSOD problem.
- SPE significantly improves the efficiency and precision of the end-to-end WSOD methods, demonstrating the potential to be a new baseline framework.

2 Related Work

2.1 Weakly Supervised Object Detection

Enumerate-and-Select Method (Two-stage). This line of methods enumerates object locations using a stand-alone region proposal algorithm. A multiple instance learning (MIL) procedure iteratively performs proposal selection and detector estimation. Nevertheless, as the object proposals are dense and redundant, MIL is often puzzled by the partial activation problem [5,49,32,13]. WSDDN [6] built the first deep MIL network by integrating an MIL loss into a deep network. Online instance classifier refinement (OICR) [18,15,44,49,54,26] was proposed to select high-quality instances as pseudo objects to refine the instance classifier. Proposal cluster learning (PCL) [24,43] further alleviated networks from concentrating on object parts by proposal clustering [43].

In the two-stage framework, object pixel gradient [39], segmentation collaboration [21,28,15,41], dissimilarity coefficient [3], attention and self-distillation [23] and extra annotations from other domains [17,7] were introduced to optimize proposal selection. Context information [25,51] was also explored to identify the instances from surrounding regions. In [49,50], a min-entropy model was proposed to alleviate localization randomness. In [26], object-aware instance labeling was explored for accurate object localization by considering the instance completeness. In [19,52], continuation MIL was proposed to alleviate the non-convexity of the WSOD loss function.

Despite the substantial progress, most WSOD methods used a stand-alone proposal generation module, which decreases not only the overall detection efficiency but also the performance upper bound.

Enumerate-and-Select Method (End-to-end). Recent methods [45,38] attempted to break the two-stage WSOD routine. WeakRPN [45] utilized object contours in convolutional feature maps to generate proposals to train a region proposal network (RPN). However, it remains relying on proposal enumeration during the training stage. In [38], an RPN [34] was trained using the pseudo objects predicted by the weakly supervised detector in a self-training fashion. Nevertheless, it requires generating dense object proposals by sliding windows. Both methods suffer from selecting inaccurate candidates from dense proposals.

2.2 Object Proposal Generation

Empirical Enumeration Method. This line of methods enumerates dense proposals based on simple features and classifiers [9,37,2]. Constrained Parametric MinCuts (CPMC) [9] produced up to 10,000 regions based on figure-ground segments and trained a regressor to select high-scored proposals. Selective Search [37] and MCG [2] adopted hierarchical segmentation and region merging on the color and contour features for proposal generation. BING [12] generated redundant proposals with sliding windows and filtered them with a classifier. EdgeBoxes [27] estimated objectness by detecting complete contours in dense bounding proposals.

Learning-based Method. Recent methods had tried to learn an RPN under weak supervision. In [45], an EdgeBoxes-like algorithm is embedded into DNNs. In [42], extra video datasets were used to learn an RPN [34]. In [38], the RPN was trained using the pseudo objects selected by the weakly supervised detector in a self-supervised fashion.

However, these methods required generating very dense object proposals. The problem of achieving a high recall rate using sparse (hundreds or tens of) proposals without precise supervision still remains.

3 Methodology

In this section, we first give an overview of the proposed sparse proposal evolution (SPE) approach. We then introduce the seed proposal generation (SPG) and sparse proposal refinement (SPR) modules. Finally, we describe the end-to-end training procedure based on iterative optimization of SPG and SPR.

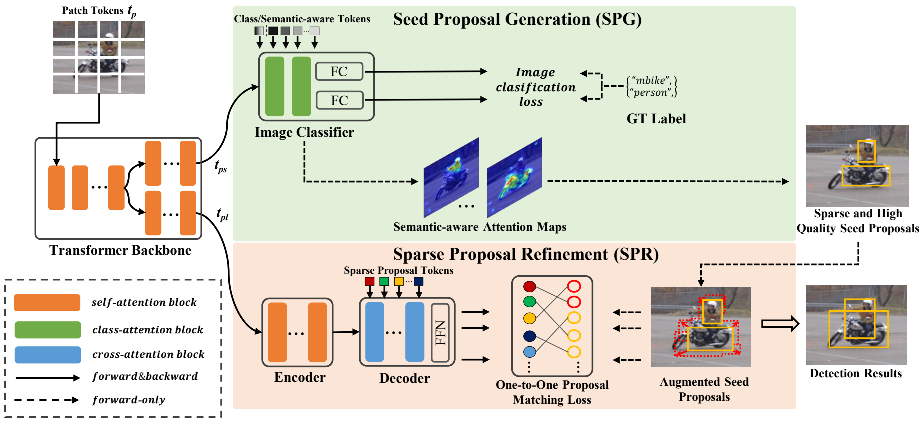


Fig. 3: Flowchart of the proposed sparse proposal evolution (SPE) approach. The diagram consists of a transformer backbone, a seed proposal generation (SPG) branch and a sparse proposal refinement (SPR) branch. During the training phase, SPG and SPR are jointly performed under a "seed-and-refine" mechanism for end-to-end WSOD with sparse object proposals.

3.1 Overview

Fig. 3 presents the flowchart of SPE, which consists of a backbone network, an SPG branch, and an SPR branch. The backbone network, which is built upon CaiT [47], contains two sub-branches with l shared transformer blocks (each block has a self-attention layer and a multi-layer perception layer). The SPG branch consists of two modules, one for image classification and the other for seed proposal generation. The initial supervisions come from the image classification loss (in the SPG branch), which drive to learn the image classifiers for semantic-aware attention maps and seed proposal generation through a thresholding algorithm [55]. The SPR branch is an encoder-decoder structure [30], which is trained by the one-to-one matching loss between seed proposals and sparse proposals. During training, an input image is first divided into $w \times h$ patches to construct $N = w \times h$ patch tokens t_p. These patch tokens are fed to the transformer to extract semantic-sensitive patch embeddings t_{ps} and location-sensitive patch embeddings t_{pl}, which are respectively fed to the SPG branch and SPR branch.

3.2 Seed Proposal Generation

The core of SPE is generating sparse yet high-quality seed proposals. Visual transformer was observed to be able to extract long-range feature dependencies by taking advantage of the cascaded self-attention mechanism, which facilitated activating and localizing full object extent [20]. This inspires us to introduce it to WSOD to produce high-quality seed proposals for object localization.

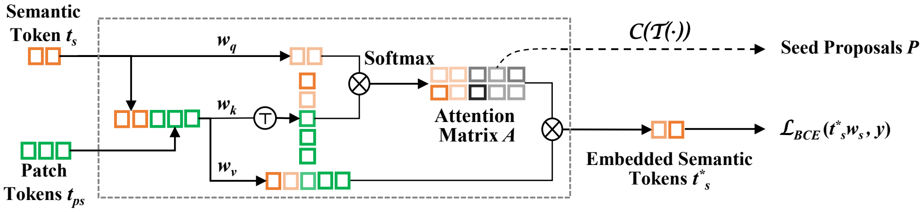


Fig. 4: Flowchart of the class-attention layer in the proposed SPG branch.

Semantic-Aware Attention Maps. As shown in Fig. 3, the SPG branch contains an image classification module and a seed proposal generation module. The image classification module contains two class-attention blocks and a fully connected (FC) layer, following CaiT [47]. Each class-attention block consists of a class-attention layer and an MLP layer with a shortcut connection. A class token $t_c \in \mathbb{R}^{1\times D}$ is fed to the first class-attention block, where the class-attention $\mathrm{CA}(\cdot)$ is performed on t_c and t_{ps} as

$$\begin{aligned} t_c^* &= \mathrm{CA}(t_c, t_{ps}, w_q, w_k, w_v) \\ &= \mathrm{Softmax}\Big((t_c w_q)([t_c, t_{ps}]w_k)^\top / \sqrt{D}\Big)([t_c, t_{ps}]w_v) \\ &= A([t_c, t_{ps}]w_v), \end{aligned} \tag{1}$$

where w_q, w_k, w_v denote weights in the class-attention layer, Fig. 4. $[t_c, t_{ps}]$ denotes concatenating t_c and t_{ps} along the first dimension. $A \in \mathbb{R}^{1\times(N+1)}$ is the attention vector of class token t_c. In the multi-head attention layer where J heads are considered, D in Eq. 1 is updated as D_0, where $D_0 = D/J$. A is then updated as the average of attention vectors weighted by their standard deviation of the J heads. t_c^* is then projected by the MLP layer in the first class-attention layer and then further fed to the second class-attention block to calculate the final embeddings $t_c^* \in \mathbb{R}^{1\times D}$ for image classification. The FC layer parameterized by $w_c \in \mathbb{R}^{D\times C}$ projects the class token t_c^* to a classification score.

Considering that the class token t_c is class-agnostic and cannot produce attention maps for each semantic category, we further add C semantic-aware tokens $t_s \in \mathbb{R}^{C\times D}$. C denotes the number of classes. By feeding both t_c and t_s to the class-attention blocks and applying class-attention defined in Eq. 1, we obtain the final token embeddings t_c^* and t_s^*. The attention vector A is updated to the attention matrix $\mathbf{A} \in \mathbb{R}^{(C+1)\times(C+N+1)}$. An extra FC layer parameterized with $w_s \in \mathbb{R}^{D\times 1}$ is added to classify the semantic tokens t_s^*. Given the image label $y = [y_1, y_2, ..., y_C]^T \in \mathbb{R}^{C\times 1}$, where $y_c = 1$ or 0 indicates the presence or absence of the c-th object category in the image, the loss function for SPG is defined as

$$\mathcal{L}_{spg}(t_c^*, t_s^*) = \mathcal{L}_{BCE}(t_c^* w_c, y) + \mathcal{L}_{BCE}(t_s^* w_s, y), \tag{2}$$

where $\mathcal{L}_{BCE}(\cdot)$ denotes the binary cross-entropy loss [6].

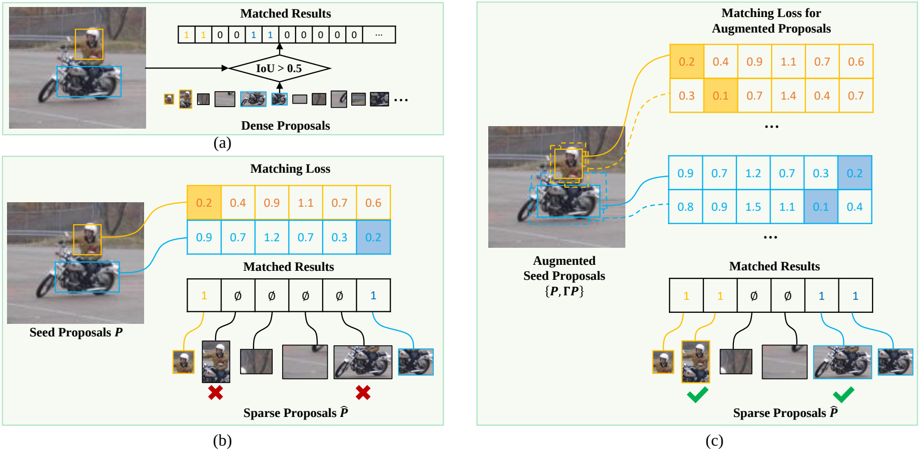


Fig. 5: Comparison of matching strategies. (a) Many-to-one matching of previous WSOD methods. (b) One-to-one matching strategy [8] applied to WSOD. (c) One-to-one matching with proposal augmentation (ours).

Seed Proposals. By optimizing Eq. 2 and executing Eq. 1 in the second class-attention block, we obtain the attention matrix $\mathbf{A} \in \mathbb{R}^{(C+1)\times(C+N+1)}$. The semantic-aware attention matrix $\mathbf{A}^* \in \mathbb{R}^{C\times N}$ is produced by indexing the first C rows and the middle N columns from $\mathbf{A}$. Attention map A_c of the c-th class is then obtained by reshaping the c-th row in $\mathbf{A}^*$ to $w \times h$ and then resized to the same resolution as the original image.

A thresholding function $\mathcal{T}(A_c, \delta_{seed})$ with a fixed threshold δ_{seed} [55] is used to binarize each semantic-aware attention map to foreground or background pixels. Based on $\mathcal{T}(A_c, \delta_{seed})$, the seed proposals are generated as

$$P = \{\mathcal{C}(\mathcal{T}(A_c, \delta_{seed}), \delta_{multi}), ...\}_{c=1}^{C} = \{B, O\} = \{(b_1, o_1), (b_2, o_2), ...\}, \qquad (3)$$

where function $\mathcal{C}(\cdot)$ outputs a set of tight bounding boxes to enclose the connected regions in the binary map $\mathcal{T}(A_c, \delta_{seed})$, under the constraint that the area of each connected region is larger than δ_{multi} of the largest connected region. Consequently, we obtain a set of bounding boxes $B = [b_1, b_2, ..., b_M] \in \mathbb{R}^{M\times 4}$ for foreground categories in the image, where each category produces at least one seed proposal. The one-hot class labels for these bounding boxes are denoted as $O = [o_1, o_2, ..., o_M]^T \in \mathbb{R}^{M\times C}$.

3.3 Sparse Proposal Refinement

Although SPG can perform object localization using seed proposals, the performance is far from satisfactory due to the lack of instance-level supervision. We further propose sparse proposal refinement (SPR), with the aim of learning object detector while refining seed proposals.

Sparse Proposals. As shown in Fig. 3, the SPR branch follows recently proposed fully-supervised transformer detectors (DETR [8] and Conditional DETR [30]), which leverage a transformer encoder, a transformer decoder, and a feed-forward network (FFN) to predict the object categories and locations. The location-sensitive patch embedding t_{pl} from the transformer backbone is first encoded by the transformer encoder to t^*_{pl}. In the transformer decoder, a fixed set of sparse proposal tokens $t_p \in \mathbb{R}^{K\times D}$ are defined to make conditional cross-attention [30] with the encoded location-aware embedding t^*_{pl}.

The decoded t^*_p is then fed to the FFN to predict K sparse proposals, as

$$\widehat{P} = \mathrm{FFN}(t^*_p, w_{FFN}) = \{\widehat{B}, \widehat{O}\} = \{(\hat{b}_1, \hat{o}_1), (\hat{b}_2, \hat{o}_2), ..., (\hat{b}_K, \hat{o}_K)\}, \tag{4}$$

where w_{FFN} and K respectively denote the parameters of the FFN and the number of proposal tokens.

One-to-One Proposal Matching. Using the seed proposals defined by Eq. 3 as pseudo objects, an optimal bipartite match between seed and sparse proposals is applied. The optimal bipartite match [8] is formulated as $\widehat{\mathfrak{S}} = [\hat{\sigma}_1, \hat{\sigma}_2, ..., \hat{\sigma}_K]$, where $\hat{\sigma}_i \in \{\varnothing, 1, 2, ..., m, ..., M\}$. $\hat{\sigma}_i = m$ denotes the i-th sparse proposal is matched with the m-th seed proposal. $\hat{\sigma}_i = \varnothing$ means that the i-th sparse proposal has no matched object and is categorized to "background". The loss function of the SPR branch is defined as

$$\begin{aligned}\mathcal{L}_{spr}(P, \widehat{P}) = \sum_{i=1}^{K} \Big[& \lambda_{FL}\mathcal{L}_{FL}(o_i, \hat{o}_{\hat{\sigma}_i}) + \mathbb{1}_{\{\hat{\sigma}_i \neq \varnothing\}}\lambda_{L_1}\mathcal{L}_{L_1}(b_i, \hat{b}_{\hat{\sigma}_i}) \\ & + \mathbb{1}_{\{\hat{\sigma}_i \neq \varnothing\}}\lambda_{GIoU}\mathcal{L}_{GIoU}(b_i, \hat{b}_{\hat{\sigma}_i}) \Big],\end{aligned} \tag{5}$$

where $\mathcal{L}_{FL}$, $\mathcal{L}_{L_1}$ and $\mathcal{L}_{GIoU}$ are Focal loss [48], L1 loss and generalized IoU loss [36], respectively. λ_{FL}, λ_{L_1} and λ_{GIoU} are regularization factors.

Seed Proposal Augmentation. The above-defined one-to-one matching breaks many-to-one label assignment, Fig. 5(a). However, the supervision signals (seed proposals) generated by attention maps contain localization noises that cannot be corrected by the one-to-one matching mechanism, Fig. 5(b). To alleviate this problem, we augment the seed proposals through a "*box jittering*" strategy, which produces randomly jittered bounding boxes on four orientations. The *box jittering* process of a bounding box $b_i = (t_x, t_y, t_w, t_h)$ is defined as

$$\mathit{\Gamma} b_i = (t_x, t_y, t_w, t_h) \pm (\varepsilon_x t_x, \varepsilon_y t_y, \varepsilon_w t_w, \varepsilon_h t_h), \tag{6}$$

where the coefficients $(\varepsilon_x, \varepsilon_y, \varepsilon_w, \varepsilon_h)$ are randomly sampled from a uniform distribution $U(-\delta_{aug}, +\delta_{aug})$. δ_{aug} is a small value to ensure $\mathit{\Gamma} b_i$ is around b_i.

By applying "*box jittering*" upon the boxes B, we extend the seed proposals $P = \{O, B\}$ to augmented seed proposals $\{P, \mathit{\Gamma} P\} = \{[O, \mathit{\Gamma} O], [B, \mathit{\Gamma} B]\}$, where the class label $\mathit{\Gamma} o_i$ is the same as o_i. With seed proposal augmentation, sparse proposals can correct noise in seed proposals, Fig. 5(c), which facilities seed proposal refinement and detection performance improvement.

Modules	δ_{seed}	δ_{multi}	mAP	CorLoc
SPG	0.1	1	23.0	48.2
	0.2	1	**29.7**	**57.8**
	0.3	1	18.2	43.6
	0.4	1	8.9	25.9
SPE	0.2	1	37.8	56.9
	0.2	0.75	41.0	61.0
	0.2	0.5	**42.6**	61.3
	0.2	0.25	42.4	**61.5**

Table 1: Performance with respect to δ_{seed} and δ_{multi} on VOC 2007 *test* set.

#SPR branches	δ_{aug}	mAP	CorLoc
0 (SPG)	0	29.7	57.8
1	0	42.6	61.3
2	0	**42.9**	**61.5**
3	0	42.7	61.3
1	0.05	42.7	61.3
1	0.1	**45.6**	**64.0**
1	0.15	45.1	64.0
1	0.2	43.4	61.6

Table 2: Performance of SPE under SPR branch numbers on VOC 2007 *test* set.

3.4 End-to-End Training

As the proposal generation and proposal refinement branches are unified upon the transformer backbone, we are able to train the seed proposal generator, the object detector, and the backbone network in an end-to-end fashion. As shown in Fig. 3, the SPG branch and the SPR branch share the transformer backbone [47]. Considering that the optimization objectives of the two network branches are not exactly the same, we separate the backbone transformer from the $(l+1)$-th block so that they share only part of the backbone network. The two network branches are jointly optimized by the total loss defined as

$$\mathcal{L}_{spe} = \mathcal{L}_{spg} + \mathbb{1}_{\{e \geq \tau\}} \mathcal{L}_{spr}, \tag{7}$$

where e denotes the training epoch and τ is a threshold number of epochs. During end-to-end training, the SPG branch is first optimized for τ epochs as a "warm-up" step, which guarantees that the seed proposals are semantic-aware and can coarsely cover object extent. Subsequently, the transformer backbone, the SPG ,and SPR branches are jointly trained under the supervision of the image classification loss and the proposal matching loss.

4 Experiment

In this section, we first introduce the experimental settings. We then conduct ablation study and quantitative and qualitative model analysis. We finally compare the proposed SPE approach with the state-of-the-art (SOTA) methods.

4.1 Experimental Setting

SPE is implemented based on the CaiT-XXS36 model [47] pre-trained on the ILSVRC 2012 dataset [1]. We evaluate SPE on the PASCAL VOC 2007, 2012 and MS COCO 2014, 2017 datasets. On VOC, we use mAP [29] and correct localization (CorLoc) [46] as the evaluation metric. The model is trained on the union set of VOC 2007 *trainval* and VOC 2012 *trainval* ("0712", containing 16551 images of 20 object classes), and evaluated on *test* set of VOC 2007 (containing 4952 images). On MS COCO datasets, we use average precision (AP) as the evaluation metric. The COCO datasets contain 80 object categories and have more challenging aspects including multi-objects and complex backgrounds. On COCO 2014 and 2017, we respectively use the 83k and 118k *train* sets for training, the 40k and 5k *val* sets for testing. Each input image is re-scaled to the fixed size and randomly horizontally flipped.

During training, we employ the AdamW gradient descent algorithm with weight decay 5e-2, and a batch size of 8 in 8 GPUs. The model respectively iterates 50 and 15 epochs on VOC and COCO datasets. During training, the learning rate for the backbone is fixed to be 1e-5. The learning rate for the rest branches is initialized to 1e-4 and drops to 1e-5 after 40 and 11 epochs on VOC and COCO datasets, respectively. The number K of proposal tokens is set to 300 following [8]. The "warm-up" time τ in Eq. 7 is empirically set to 7.

4.2 Ablation Study

We analyze SPE's hyper-parameters δ_{seed} and δ_{multi}, times of proposal refinement, matching manners, and the detection efficiency. We also study the effect of the shared backbone block numbers, the detector and backbone network. All the ablation experiments are conducted on PASCAL VOC.

SPG. Table 1 includes the detection and localization performance of SPG under different δ_{seed} and δ_{multi}. It can be seen that δ_{seed} has the key influence from the generation of seed proposals. When $\delta_{seed} = 0.2$, SPG achieves 29.7% mAP and 57.8% CorLoc. When δ_{multi} decreases, the performance first increases and then decreases. This implies that as δ_{multi} decreases SPG discovers more and more objects, which enriches the supervision signals and improves the detection performance. On the other hand, with the increase of δ_{multi}, SPG produces more noise proposals, which degenerate the detection performance.

SPR. Table 2 shows the effect of proposal refinement times by adding extra SPR branches and introducing seed proposal augmentation. When adding one SPR branch, the detection performance is significantly improved by 12.9%(29.7% vs 42.6%) and the localization performance is improved by 3.5%(57.8% vs 61.3%), which clearly demonstrates SPR's effectiveness for refining the seed proposals. When more SPR branches are added, marginal performance improvements are achieved. By introducing seed proposal augmentation, the performance is further significantly improved by 3.0%(42.6 vs 45.6%) and 2.7%(61.3 vs 64.0%) with $\delta_{aug} = 0.1$, demonstrating that the proposal augmentation mechanism can suppress the noise of seed proposals and achieve more accurate localization.

Detection Efficiency. In Table 3, we compare the proposed SPE with "enumerate-and-select" methods, including the two-stage OICR method [44] and the STOA end-to-end method UWSOD [38]. The compared terms include the number of parameters (#Params), MACs, time of proposal generation (τ) and inference speed. The experiments are carried out under image scale 512^2 on the PASCAL VOC dataset (0712 *trainval* for training and 07 *test* set for testing).

Methods	#Params (M)	MACs (G)	τ (s/img)	speed(fps)	mAP
OICR(VGG16) [44]	120.9	304.26	3.79	0.26	44.1
UWSOD(VGG16) [38]	138.5	923.31	0.002	4.2	45.7
UWSOD(WSR18) [38]	135.0	237.97	0.002	4.3	46.9
SPE(Ours)	33.9	51.25	0	14.3	51.0

Table 3: Comparison of parameters, MACs, time to generate proposals and inference speed on the VOC *test* set. SPE is implemented based on CaiT-XXS36. Test speeds ("speed") are evaluated on a single NVIDIA RTX GPU.

l shared blocks	mAP	CorLoc
36	32.8	50.0
24	**45.6**	**64.0**
12	43.1	60.1

Table 4: Performance of SPE with l shared blocks on VOC 2007 *test* set.

Detector	Backbone	mAP
Faster RCNN	VGG16	78.3
Faster RCNN	ResNet50	80.9
Faster RCNN	CaiT-XXS36	81.4
Conditional DETR	CaiT-XXS36	77.5

Table 5: Performance of different detectors on VOC 2007 *test* set.

SPE has much fewer parameters than OICR and UWSOD (only ∼1/4 of OICR and UWSOD), and uses much fewer MACs than OICR and UWSOD (only 1/20∼1/4 of OICR and UWSOD). These results show that SPE, which discards dense proposals by learning sparse proposals, is efficient for object detection. For testing, SPE directly uses the backbone and the SPR branch for object detection and does not need computational costs for proposal generation. With such high detection efficiency, SPE achieves 51.0% mAP, which respectively outperforms OICR and UWSOD by 9.8% and 7.0%.

Number of Shared Backbone Blocks. We analyze the effect of backbone blocks shared by SPG and SPR (denoted by l), Table 4. When $l = 36$, *i.e.*, the two branches share all backbone blocks of CaiT-XXS36, the detection and localization performance are 32.8% and 50.0%, respectively. This is because the learning of regression task will interfere the attention map in SPG, and thus degenerates the quality of generated seed proposals. When $l = 24$, the above problem is largely alleviated, and the detection and localization performances respectively increase to 45.6% and 64.0%. When sharing fewer layers, the performances slightly decrease due to the increase of inductive bias.

Backbone and Detector. In Table 5, we compare Faster RCNN w/ VGG16, Faster RCNN w/ ResNet50, Faster RCNN w/ CaiT-XXS36, and Conditional DETR w/ CaiT-XXS36 on VOC 0712 under fully supervised settings. The mAP of Conditional DETR w/ CaiT-XXS36 is 77.5%, which is lower than that of Faster-RCNN w/ VGG16 (78.3%). It shows the detector is not the key factor of performance gain. The mAP of Faster R-CNN w/ CaiT-XXS36 is 81.4%, which is 3.1% higher than w/ VGG16 and 0.5% higher than w/ ResNet50. We also conducted experiments of MIST [35] w/ CaiT-XXS36, but achieved much worse results than MIST [35] w/ VGG16. Although CaiT-XXS36 is better in fully su-

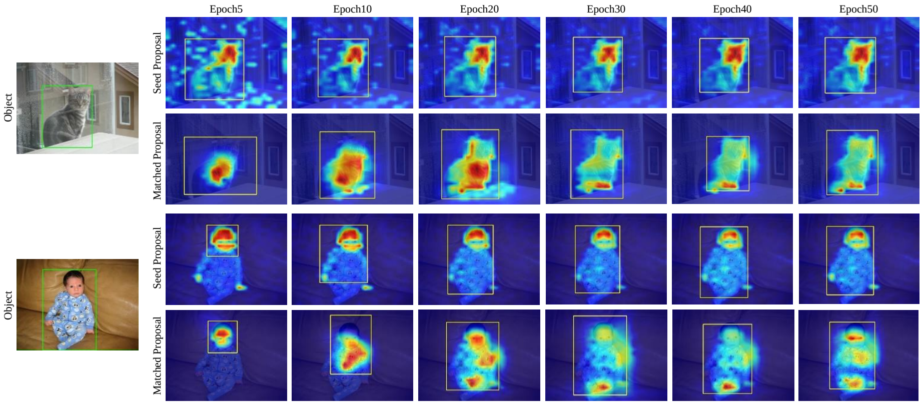


Fig. 7: Evolution of seed proposals and matched sparse proposals (yellow bounding boxes) during training. Heatmaps in the "seed proposal" column show the semantic-aware attention maps, while heatmaps in "matched proposal" column show the cross-attention maps of the matched sparse proposals.

pervised detection task, it is not superior than VGG16 for traditional WSOD methods.

4.3 Visualization Analysis

Qualitative Analysis. Fig. 7 shows the evolution of seed proposals and matched proposals and their corresponding attention maps (heatmaps) generated by SPG and SPR. At early training epochs, SPG activates most of the objects and can produce seed proposals for object location initialization. However, these proposals still suffer from background activation or partial activation. After matching with the sparse proposals, seed proposals are refined to more accurate object locations, which demonstrates the effectiveness of the SPR module with the proposal augmentation strategy. As training goes on, the seed proposals are gradually refined by and matched with the sparse proposals, and finally evolve to full object extent.

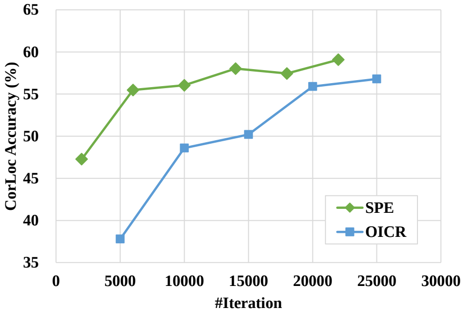


Fig. 6: Comparison of CorLoc accuracy of SPE and OICR [44] during training.

Fig. 8 visualizes the seed proposals and matched sparse proposals and the corresponding attention map. With the long-range feature dependencies of trans-

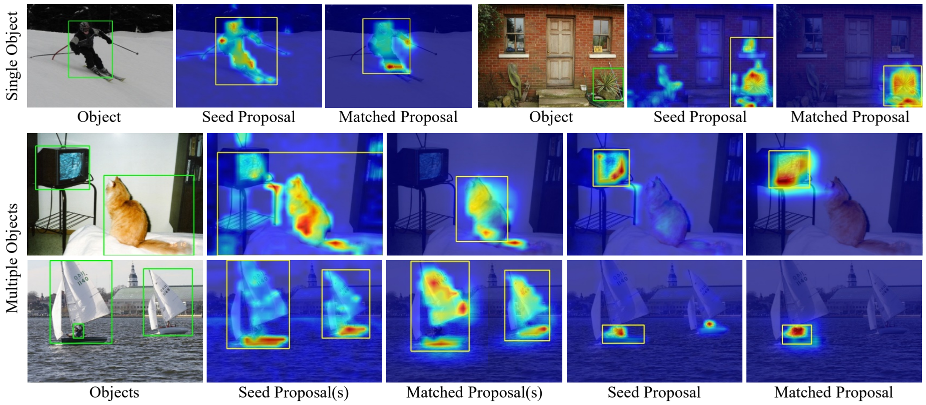


Fig. 8: Visualization of seed proposals and matched sparse proposals (yellow boxes). Heatmaps in "seed proposal" column show the semantic-aware attention maps for object classes. Heatmaps in the "matched proposal" column show the cross-attention maps of the matched sparse proposals.

former, the semantic-aware attention maps in SPG can activate full object extent. Based on these attention maps, SPG can generate sparse yet high-quality seed proposals. By introducing SPR, the matched proposal can promote seed proposals and achieves preciser object localization. These results validate the effectiveness of the proposed SPG and SPR branches of SPE, where the seed proposals and sparse proposals evolve towards true object locations.

Quantitative Analysis. Fig. 6 shows the CorLoc accuracy of SPE and OICR [44] during training iterations. By introducing transformer block, SPE can generate much preciser proposals at very early iterations. In contrast, OICR suffers from dense and noise proposals, which struggles to select the object proposals so that the localization performance deteriorates in early iterations.

4.4 Performance

PASCAL VOC. Table 6 shows the performance of SPE and the SOTA methods on VOC 2007 dataset. "07" in "Set" column denotes the *trainval* set of VOC 2007, "0712" denotes *trainval* set of VOC 2007 and 2012 datasets. "CaiT" denotes CaiT-XXS36. † refers to our implementation using the official code. With image scale 384, SPE achieves competitive 48.5% mAP and 66.4% CorLoc accuracy when training on 0712 *trainval* set. With image scale 512, SPE achieves 51.0% mAP and 70.4% CorLoc accuracy, which outperforms the two-stage methods WSDDN [6] and OICR [44] by 16.2% and 9.8%. Compared with the end-to-end methods using dense object proposals, the performance of SPE is very competitive. It also outperforms the SOTA UWSOD by 7.0% mAP.

MS COCO. In Table 7, we report the performance of SPE and the SOTA methods on the MS COCO 2014 and 2017 datasets. "14" and "17" in "Set"

Backbone	Set	Method	mAP	CorLoc
Enumerate-and-Select Methods (Two-Stage)				
VGG16	07	WSDDN [6]	34.8	53.5
	07	OICR [44]	41.2	60.6
	07	SLV [10]	53.5	**71.0**
	07	DC-WSOD [3]	52.9	70.9
	07	TS^{2}C [51]	44.3	61.0
	07	SDCN [28]	50.2	68.6
	07	C-MIL [19]	50.5	65.0
	07	PCL [43]	43.5	62.7
	07	MIST [35]	**54.9**	68.8
	0712	WSDDN† [6]	36.9	56.8
	0712	OICR† [44]	43.6	61.7
Enumerate-and-Select Methods (End-to-End)				
VGG16	07	OM+MIL [18]	23.4	41.2
	07	OPG [40]	28.8	43.5
	07	SPAM [22]	27.5	-
	07	UWSOD [38]	44.0	63.0
Seed-and-Refine Methods (End-to-End)				
CaiT	0712	SPE (ours)-384	48.5	66.4
	0712	SPE (ours)-512	**51.0**	**70.4**

Table 6: Detection Performance(%) on the PASCAL VOC 2007 *test* set.

Backbone	Set	Method	AP	AP50	AP75
Enumerate-and-Select Methods (Two-Stage)					
VGG16	14	WSDDN [6]	-	11.5	-
	14	WCCN [15]	-	12.3	-
	14	ODGA [16]	-	12.8	-
	14	PCL [43]	8.5	19.4	-
	14	WSOD^2 [54]	10.8	22.7	-
	14	C-MIDN [21]	9.6	21.4	-
	14	MIST [35]	**12.4**	**25.8**	**10.5**
	14	PG-PS [11]	-	20.7	-
Enumerate-and-Select Methods (End-to-End)					
VGG16	17	UWSOD [38]	2.5	9.3	1.1
WSR18	17	UWSOD [38]	3.1	10.1	1.4
Seed-and-Refine Methods (End-to-End)					
CaiT	14	SPE (ours)-384	5.7	15.2	3.4
	17	SPE (ours)-384	6.3	16.3	4.0
	17	SPE (ours)-512	**7.2**	**18.2**	**4.8**

Table 7: Detection Performance(%) on the MS COCO 2014 and 2017 set.

column respectively denote training on MS COCO 2014 and 2017 datasets. On COCO 2014, SPE respectively achieves 5.7%, 15.2%, and 3.4% under metric AP, AP50 and AP75, which are comparable with the two-stage "enumerate-and-select" methods. On MS COCO 2017, SPE respectively achieves 6.3% AP, 16.3% AP50 and 4.0% AP75, outperforming the end-to-end UWSOD method [38] by 3.2%, 6.2% and 2.6%. When increasing the image scale to 512, the APs are further improved by 0.9%, 1.9%, and 0.8%, respectively.

5 Conclusion

We proposed the sparse proposal evolution (SPE) approach, and advanced WSOD methods with dense proposals to an end-to-end fashion with sparse proposals. SPE uses a "seed-and-refine" framework, which is efficient for both training and test. By taking advantage of the visual transformer, SPE generates sparse yet high-quality seed proposals. With the one-to-one proposal matching strategy, SPE iteratively improves seed proposals and object detectors in a self-evolution fashion. As the first end-to-end framework with sparse proposals, SPE demonstrates tremendous potential and provides a fresh insight to the challenging WSOD problem.

Acknowledgement: This work was supported by National Natural Science Foundation of China (NSFC) under Grant 62006216, 61836012, 62171431 and 62176260, the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDA27000000.

References

1. Alex, K., Ilya, S., E, H.G.: Imagenet classification with deep convolutional neural networks. In: NeurIPS. pp. 1097–1105 (2012) 9
2. Arbeláez, P.A., Pont-Tuset, J., Barron, J.T., Marqués, F., Malik, J.: Multiscale combinatorial grouping. In: IEEE CVPR. pp. 328–335 (2014) 4
3. Arun, A., Jawahar, C.V., Kumar, M.P.: Dissimilarity coefficient based weakly supervised object detection. In: IEEE CVPR. pp. 9432–9441 (2019) 3, 14
4. Bilen, H., Pedersoli, M., Tuytelaars, T.: Weakly supervised object detection with posterior regularization. In: BMVC. pp. 1997–2005 (2014) 1
5. Bilen, H., Pedersoli, M., Tuytelaars, T.: Weakly supervised object detection with convex clustering. In: IEEE CVPR. pp. 1081–1089 (2015) 3
6. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: IEEE CVPR. pp. 2846–2854 (2016) 2, 3, 6, 13, 14
7. Cao, T., Du, L., Zhang, X., Chen, S., Zhang, Y., Wang, Y.: Cat: Weakly supervised object detection with category transfer (2021) 3
8. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229. Springer (2020) 7, 8, 10
9. Carreira, J., Sminchisescu, C.: CPMC: automatic object segmentation using constrained parametric min-cuts. IEEE TPAMI **34**(7), 1312–1328 (2012) 4
10. Chen, Z., Fu, Z., Jiang, R., Chen, Y., Hua, X.: SLV: spatial likelihood voting for weakly supervised object detection. In: IEEE CVPR. pp. 12992–13001 (2020) 14
11. Cheng, G., Yang, J., Gao, D., Guo, L., Han, J.: High-quality proposals for weakly supervised object detection. IEEE TIP **29**, 5794–5804 (2020) 14
12. Cheng, M., Zhang, Z., Lin, W., Torr, P.H.S.: BING: binarized normed gradients for objectness estimation at 300fps. In: IEEE CVPR. pp. 3286–3293 (2014) 4
13. Chong, W., Kaiqi, H., Weiqiang, R., Junge, Z., Steve, M.: Large-scale weakly supervised object localization via latent category learning. IEEE TIP **24**(4), 1371–1385 (2015) 3
14. Chong, W., Weiqiang, R., Kaiqi, H., Tieniu, T.: Weakly supervised object localization with latent category learning. In: ECCV. pp. 431–445 (2014) 1
15. Diba, A., Sharma, V., Pazandeh, A., Pirsiavash, H., Van Gool, L.: Weakly supervised cascaded convolutional networks. In: IEEE CVPR. pp. 5131–5139 (2017) 3, 14
16. Diba, A., Sharma, V., Stiefelhagen, R., Van Gool, L.: Object discovery by generative adversarial & ranking networks. arXiv preprint arXiv:1711.08174 (2017) 14
17. Dong, B., Huang, Z., Guo, Y., Wang, Q., Niu, Z., Zuo, W.: Boosting weakly supervised object detection via learning bounding box adjusters. In: IEEE ICCV (2021) 3
18. Dong, L., Bin, H.J., Yali, L., Shengjin, W., Hsuan, Y.M.: Weakly supervised object localization with progressive domain adaptation. In: IEEE CVPR. pp. 3512–3520 (2016) 3, 14
19. Fang, W., Chang, L., Wei, K., Xiangyang, J., Jianbin, J., Qixiang, Y.: Cmil: Continuation multiple instance learning for weakly supervised object detection. In: IEEE CVPR (2019) 2, 4, 14
20. 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. CoRR **abs/2103.14862** (2021) 5

21. Gao, Y., Liu, B., Guo, N., Ye, X., Wan, F., You, H., Fan, D.: C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In: IEEE ICCV (2019) 3, 14
22. Gudi, A., van Rosmalen, N., Loog, M., van Gemert, J.C.: Object-extent pooling for weakly supervised single-shot localization. In: BMVC (2017) 14
23. Huang, Z., Zou, Y., Kumar, B.V.K.V., Huang, D.: Comprehensive attention self-distillation for weakly-supervised object detection. In: NeurIPS (2020) 2, 3
24. Jie, Z., Wei, Y., Jin, X., Feng, J., Liu, W.: Deep self-taught learning for weakly supervised object localization. In: IEEE CVPR. pp. 4294–4302 (2017) 3
25. Kantorov, V., Oquab, M., Cho, M., Laptev, I.: Contextlocnet: Context-aware deep network models for weakly supervised localization. In: ECCV. pp. 350–365 (2016) 4
26. Kosugi, S., Yamasaki, T., Aizawa, K.: Object-aware instance labeling for weakly supervised object detection. In: IEEE ICCV (2019) 3, 4
27. Lawrence, Z.C., Piotr, D.: Edge boxes: Locating object proposals from edges. In: ECCV. pp. 391–405 (2014) 2, 4
28. Li, X., Kan, M., Shan, S., Chen, X.: Weakly supervised object detection with segmentation collaboration. In: IEEE ICCV (2019) 3, 14
29. Mark, E., Luc, V.G., KI, W.C., John, W., Andrew, Z.: The pascal visual object classes (voc) challenge. IJCV **88**(2), 303–338 (2010) 9
30. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional detr for fast training convergence. In: IEEE ICCV. pp. 3651–3660 (October 2021) 5, 8
31. Oh, S.H., Jae, L.Y., Stefanie, J., Trevor, D.: Weakly supervised discovery of visual pattern configurations. In: NeurIPS. pp. 1637–1645 (2014) 1
32. Oh, S.H., Ross, G., Stefanie, J., Julien, M., Zaid, H., Trevor, D.: On learning to localize objects with minimal supervision. In: ICML. pp. 1611–1619 (2014) 1, 3
33. Parthipan, S., Tao, X.: Weakly supervised object detector learning with model drift detection. In: IEEE ICCV. pp. 343–350 (2011) 1
34. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS. pp. 91–99 (2015) 4
35. Ren, Z., Yu, Z., Yang, X., Liu, M., Lee, Y.J., Schwing, A.G., Kautz, J.: Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In: IEEE CVPR. pp. 10595–10604 (2020) 11, 14
36. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: IEEE CVPR (June 2019) 8
37. RR, U.J., de Sande Koen EA, V., Theo, G., WM, S.A.: Selective search for object recognition. IJCV **104**(2), 154–171 (2013) 2, 4
38. Shen, Y., Ji, R., Chen, Z., Wu, Y., Huang, F.: UWSOD: toward fully-supervised-level capacity weakly supervised object detection. In: NeurIPS (2020) 2, 4, 10, 11, 14
39. Shen, Y., Ji, R., Wang, C., Li, X., Li, X.: Weakly supervised object detection via object-specific pixel gradient. IEEE TNNLS **29**(12), 5960–5970 (2018) 3
40. Shen, Y., Ji, R., Wang, C., Li, X., Li, X.: Weakly supervised object detection via object-specific pixel gradient. IEEE TNNLS **29**(12), 5960–5970 (2018) 14
41. Shen, Y., Ji, R., Wang, Y., Wu, Y., Cao, L.: Cyclic guidance for weakly supervised joint detection and segmentation. In: IEEE CVPR. pp. 697–707 (2019) 3
42. Singh, K.K., Lee, Y.J.: You reap what you sow: Using videos to generate high precision object proposals for weakly-supervised object detection. In: IEEE CVPR. pp. 9414–9422 (2019) 2, 4

43. 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 **42**(1), 176 – 191 (2020) 3, 14
44. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: IEEE CVPR. pp. 3059–3067 (2017) 2, 3, 10, 11, 12, 13, 14
45. 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) 2, 4
46. Thomas, D., Bogdan, A., Vittorio, F.: Weakly supervised localization and learning with generic knowledge. IJCV **100**(3), 275–293 (2012) 9
47. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. arXiv preprint arXiv:2103.17239 (2021) 2, 5, 6, 9
48. Tsung-Yi, L., Priya, G., Ross, G., Kaiming, H., Dollár, P.: Focal loss for dense object detection. In: IEEE ICCV (2017) 8
49. Wan, F., Wei, P., Jiao, J., Han, Z., Ye, Q.: Min-entropy latent model for weakly supervised object detection. In: IEEE CVPR. pp. 1297–1306 (2018) 3, 4
50. Wan, F., Wei, P., Jiao, J., Han, Z., Ye, Q.: Min-entropy latent model for weakly supervised object detection. IEEE TPAMI **41**(10), 2395 – 2409 (2019) 4
51. Wei, Y., Shen, Z., Cheng, B., Shi, H., Xiong, J., Feng, J., Huang, T.: Ts2c:tight box mining with surrounding segmentation context for weakly supervised object detection. In: ECCV. pp. 434–450 (2018) 4, 14
52. Ye, Q., Wan, F., Liu, C., Huang, Q., Ji, X.: Continuation multiple instance learning for weakly and fully supervised object detection. IEEE TNNLS pp. 1–15 (2021). https://doi.org/10.1109/TNNLS.2021.3070801 2, 4
53. Ye, Q., Zhang, T., Qiu, Q., Zhang, B., Chen, J., Sapiro, G.: Self-learning scene-specific pedestrian detectors using a progressive latent model. In: IEEE CVPR. pp. 2057–2066 (2017) 1
54. 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 (2019) 3, 14
55. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: IEEE CVPR. pp. 2921–2929 (2016) 5, 7