Extreme Point Supervised Instance Segmentation

Hyeonjun Lee[1,3]   Sehyun Hwang[2]   Suha Kwak[2,3]

[1]Lunit Inc.   [2]Dept. of CSE, POSTECH   [3]Graduate School of AI, POSTECH

dlguswns1882@gmail.com, {sehyun03, suha.kwak}@postech.ac.kr

Abstract

This paper introduces a novel approach to learning instance segmentation using extreme points, i.e., the topmost, leftmost, bottommost, and rightmost points, of each object. These points are readily available in the modern bounding box annotation process while offering strong clues for precise segmentation, and thus allows to improve performance at the same annotation cost with box-supervised methods. Our work considers extreme points as a part of the true instance mask and propagates them to identify potential foreground and background points, which are all together used for training a pseudo label generator. Then pseudo labels given by the generator are in turn used for supervised learning of our final model. On three public benchmarks, our method significantly outperforms existing box-supervised methods, further narrowing the gap with its fully supervised counterpart. In particular, our model generates high-quality masks when a target object is separated into multiple parts, where previous box-supervised methods often fail.

1. Introduction

Instance segmentation, the task of predicting classes and masks of individual objects at the same time, has been advanced remarkably thanks to supervised learning of deep neural networks [9, 21, 56, 58, 59]. However, it is prohibitively costly to manually annotate a pixel-level mask per instance, which often leads to lack of both class diversity and the amount of training data. This issue steers the research community towards label-efficient learning approaches such as weakly supervised learning [1, 10, 13, 23, 27, 28, 34–36, 38, 39, 54, 57, 70] and semi-supervised learning [24, 26, 29, 42, 47, 52, 61, 69].

Building on this momentum, learning instance segmentation using box supervision has gained considerable attraction recently [13, 23, 27, 34–36, 38, 39, 57]. To train an instance segmentation model with box-supervision, these methods employ a bounding box tightness prior [23], which implies that a vertical (or horizontal) line crossing the bounding box must contain at least one pixel belonging to the object

(Fig. 1); this prior has been formulated through various loss functions [13, 23, 34, 35, 39, 57]. Although box-supervision has proved to be effective for learning instance segmentation while keeping annotation costs low, we claim that there is room for further improvement in this direction, particularly due to the fact that it has neglected *extreme points*, a byproduct of the common box annotation process providing a strong clue that helps in estimating the instance mask.

Today, extreme points are freely available in the bounding box annotation process [32], where human annotators are instructed to click four extreme points of the target object, *i.e.*, topmost, leftmost, bottommost, and rightmost points, rather than to click two corner points of the bounding box. This is because the former usually ends up requiring less annotation time as the latter often needs to adjust the initial box label multiple times, as demonstrated by Papadopoulos *et al.* [48]. Moreover, since they are definitely a part of the true mask of the target, extreme points provide a strong clue for segmentation absent in the box supervision.

Motivated by this, we study weakly supervised learning for instance segmentation using extreme points to further improve performance without increasing annotation cost. Our framework for EXtreme point supervised InsTance Segmentation, dubbed EXITS, considers extreme points as a part of the true instance mask, and exploits them as supervision for training a pseudo label generator. Then pseudo segmentation labels produced by the generator are in turn used for supervised learning of our final model, which can be any arbitrary networks for instance segmentation. The overall procedure of EXITS is illustrated in Fig. 2.

The key to the success of EXITS is how to train the pseudo label generator using extreme points. A straightforward way is to consider extreme points as foreground and points outside the bounding box as background, and then exploit them for supervised learning. However, the pseudo label generator trained in this way fails to generate crisp object masks since most object regions remain unlabeled during training due to the sparsity of extreme points. To address this issue, EXITS estimates potential foreground and background points within the bounding box by propagating the extreme and background points outside the box. The propagation
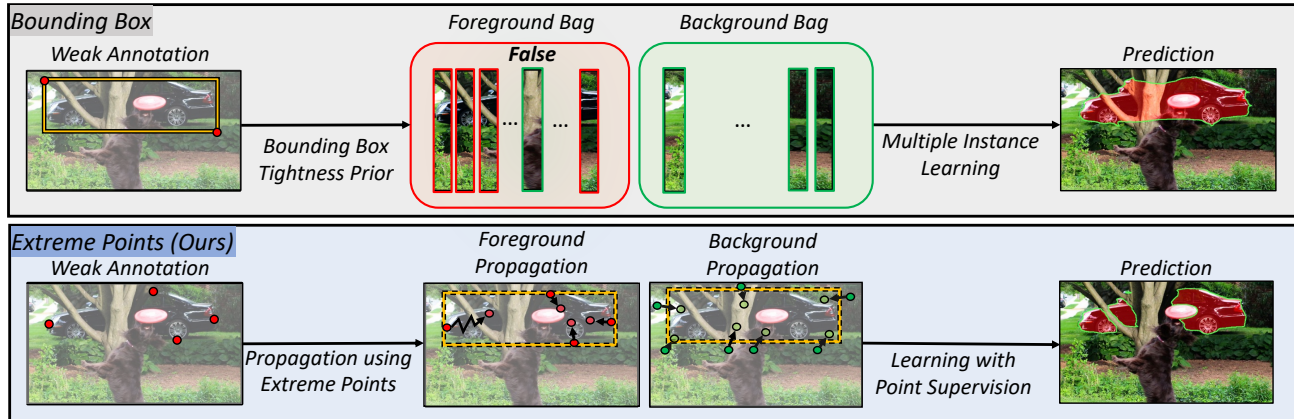
Figure 1. Types of weak supervision and how to utilize it for instance segmentation. Top: Box-supervised method relies on bounding box tightness prior, which is often violated by occlusion (foreground bag contains tree trunk). As a result, the prediction of the method shows an error in the occluded region. Bottom: Extreme point supervised method (Ours) utilizes extreme points as the initial set of foreground points and propagate label through semantic similarity between points. The prediction result demonstrates that our method can predict object mask even in severe occlusion. Best viewed in color.

process is based on pairwise semantic similarity between points derived by a pretrained transformer encoder so that it reveals foreground and background candidates semantically similar with extreme points and nearby background, respectively. The retrieved points together with the extreme and definite background points serve as supervision for training the pseudo label generator.

As shown in Fig. 1, our pseudo label generator produces high-quality pseudo masks, particularly when a target is divided into multiple parts, and the enhanced quality of pseudo segmentation labels leads to performance improvement of our final model. This success is due to the fact that the label propagation is conducted on the fully connected graphs of all the points so that an extreme point can be propagated to spatially distant points. This alleviates the side-effect of the bounding box tightness prior that is violated in the case of occlusion; the convention box-supervised methods, which rely heavily on the prior, thus often failed in the case.

To quantitatively compare the quality of pseudo labels for separated objects, we measured the pseudo label quality on Separated COCO [64], a subset of COCO [40] comprising only separated objects. On the dataset, our method surpassed the previous best method [35] by 7.3%p in mIoU. We further evaluated EXITS on three public benchmarks, PASCAL VOC [17], COCO, and LVIS [18], where EXITS outperformed all the previous box-supervised methods.

In short, the main contribution of this paper is three-fold:

- We tackle weakly supervised instance segmentation using extreme points, which can be obtained during bounding box labeling without extra costs.
- We introduce a point retrieval algorithm, which effectively leverages extreme points to estimate labels of points in the bounding box. Specifically, this algorithm estimates the

labels of points based on the probability of propagation to extreme points and background points.
- Our Method achieved the state of the art on three public benchmarks. The qualitative results demonstrated that our method generates high-quality pseudo masks, particularly for separated objects.

2. Related Work

Instance segmentation. Mask R-CNN [21] proposes a two-stage approach that first detects regions of interest (RoI) and then predicts segmentation masks within these RoIs. Subsequent studies have refined this concept by enhancing feature representation [4, 6, 41] or mask precision [11, 25, 65]. Then, one-stage methods [3, 12, 56, 62, 66] typically built upon one-stage detectors [49, 55] have gained attractions, thanks to their speed and simplicity. Meanwhile, methods like SOLO [58] and SOLOv2 [59] introduce box-free one-stage methods without the need for box prediction. Recently, query-based methods [8, 9, 14, 19, 37], inspired by DETR [5], offer impressive performance. Although these fully supervised methods show remarkable performance, they face practical challenges due to their dependence on costly pixel-wise mask annotation.

Weakly supervised instance segmentation. Weakly supervised methods using image-level class labels [1, 28, 68, 70], which depends heavily on class activation maps, have not yet matched fully-supervised performance. Box-supervised methods offer better results with lower annotation costs. The first method in this direction [27] refines pseudo masks using GrabCut [51], while recent methods [34, 39, 57, 60] incorporate bounding box tightness priors and Multiple Instance Learning (MIL) loss, enhanced with techniques like saliency, color-pairwise affinity, and semantic-correspondence. An-

(1) Learning pseudo label generator with extreme point

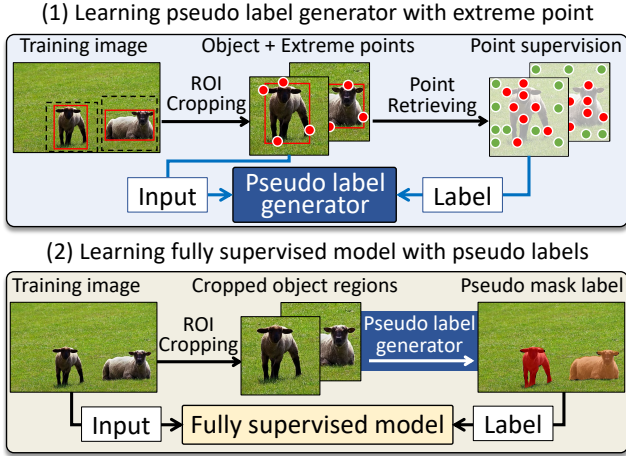(2) Learning fully supervised model with pseudo labels

Figure 2. Overview of entire stages of EXITS. In the first stage, an image cropped around each object is used as an input to train the pseudo label generator using point-wise supervision, so that the generator learns to predict a binary mask of the object within the cropped image. In the second stage, the instance segmentation model learns to detect and segment multiple objects, using the generated pseudo mask labels from the first stage.

other trend includes the Mask Auto Labeler (MAL) [35], which uses a two-stage process involving pseudo mask generation and model training. Point-based methods [10, 54] add point labels to boxes for improved localization. In contrast, our approach leverages extreme points obtained from box annotations for weak supervision, offering robust clues for instance mask estimation.

Extreme point for object annotation. An extreme point label is an efficient alternative to a bounding box label, offering a faster annotation process [48]. This approach, being five times quicker than traditional methods, has been increasingly used in object detection training [32, 67] and object segmentation tasks [15, 46, 48, 50]. DEXTR [46], for instance, utilizes extreme points for segmenting arbitrary objects by learning the mapping between input images with extreme points and their segmentation masks. However, DEXTR still requires expensive pixel-level masks for training. In medical imaging, methods like [15, 50] use extreme points for training voxel segmentation models, generating pseudo-scribble labels by linking extreme points via the shortest path. Despite these benefits of extreme point label, it has received limited attention in weakly-supervised instance segmentation. Motivated by this, we introduce to leverage extreme point labels for instance segmentation in diverse scenes predicting precise object masks without using pixel-wise annotations. Unlike typical approaches in medical imaging that generate scribble pseudo labels based on path-connected object regions, our method uses extreme points to select pseudo-foreground points, which is crucial in scenarios with occlusions, as demonstrated in Fig. 1.

3. Proposed Method

EXITS consists of two stages: (1) learning a model that generates pseudo segmentation labels of training images using their extreme point labels, and (2) training an instance segmentation model using the pseudo labels. In the first stage, an object image cropped around each object using extreme points is used as an input to the pseudo label generator so that the model learns to predict a binary mask of the object within the cropped image. On the other hand, the instance segmentation model in the second stage, which is our final model, learns to detect and segment multiple objects. Note that the pseudo label generator deals with an easier task, *i.e.*, instance segmentation on a single object image, which enables to improve the quality of pseudo labels it generates. The entire pipeline of EXITS is illustrated in Fig. 2.

Since the second stage is the conventional supervised learning that can be applied to any instance segmentation model, this section elaborates mostly on the first stage, in particular, how EXITS provides the pseudo label generator with effective supervision learning for segmentation. The overall pipeline of the first stage is illustrated in Fig. 3. The key idea of EXITS is to retrieve pixels likely to belong to the object given the extreme points, and exploit them as supervision for the pseudo label generator. This idea is realized by propagating the extreme points to other pixels within the input object image, while considering the extreme points as a subset of true pixels of the object.

The remainder of this section first discusses extreme points and advantages of using them (Sec. 3.1), and then presents details of the pseudo label generator (Sec. 3.2) and the second stage (Sec. 3.3) of EXITS.

3.1. Motivation for Using Extreme Points

Extreme points are defined as the outermost pixels on an object along the cardinal directions: the topmost point $(x^{(t)}, y^{(t)})$, the leftmost point $(x^{(l)}, y^{(l)})$, the bottommost point $(x^{(b)}, y^{(b)})$, rightmost point $(x^{(r)}, y^{(r)})$. Papadopoulos *et al.* [48] demonstrated labeling these points is a more efficient way to bounding box annotation compared to the conventional method of labeling the top-left $(x^{(l)}, y^{(t)})$ and bottom-right $(x^{(r)}, y^{(b)})$ corner points of a box. This is because such corner points are hard to be identified as they usually do not belong to the target object area, and thus human annotators often have to adjust their initial corner point labels several times. On the other hand, extreme points can be effortlessly marked and directly converted to a bounding box. Furthermore, they inherently provide more information for the shape and appearance of the target object than corner points since they lie on the object boundary.

3.2. Learning Pseudo Label Generator

The pseudo label generator aims to predict a binary mask of an object given an image cropped around it. It consists
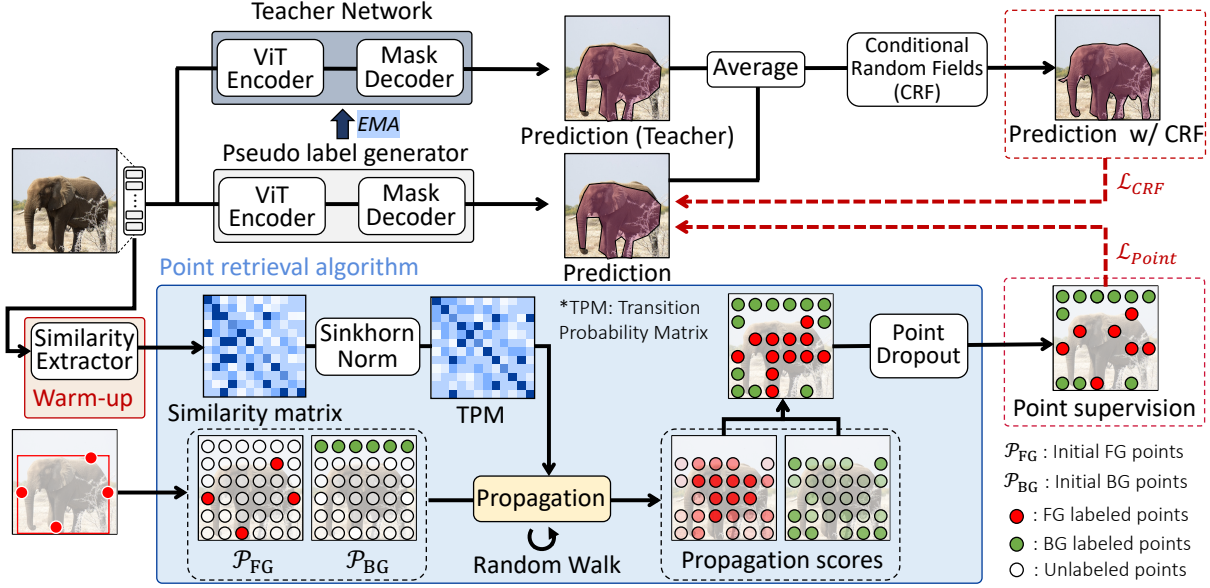
Figure 3. Overview of the first stage of EXITS framework. The pseudo label generator is trained on images cropped around each object using the extreme points, aiming to predict binary masks. Training leverages two loss functions: $\mathcal{L}_{crf}$ aligns images before and after CRF [33] processing, and $\mathcal{L}_{point}$ uses extreme points-derived pseudo point labels for precise pixel-wise supervision. To generate these pseudo point labels, EXITS obtains initial foreground and background points from extreme points, then employs the similarity matrix from warm-up trained similarity extractor for label propagation. After propagation, pseudo point labels are produced based on the difference of propagation score from the inital foreground and background points. Point dropout is applied as an augmentation generating the final pseudo point labels.

of a vision transformer (ViT) encoder and a mask decoder. We retrieve points likely to belong to the object (*i.e.*, foreground) or the background, and train the generator using the retrieved points together with the extreme points and definite background points outside the box as supervision.

To be specific, the initial set of foreground points is derived from the extreme points as $\mathcal{P}_{FG} := \{(x^{(t)}, y^{(t)} - \delta), (x^{(l)} + \delta, y^{(l)}), (x^{(b)}, y^{(b)} + \delta), (x^{(r)} - \delta, y^{(r)})\}$, where δ is a small margin introduced to push the extreme points toward the center of the object so that the points in $\mathcal{P}_{FG}$ are more inward and represent the object more reliably. On the other hand, the initial set of background points $\mathcal{P}_{BG}$ consists of points located outside the bounding box defined by the extreme points. To assign pseudo labels to unlabeled points within the bounding box, denoted as $\mathcal{P}_{Box}$, the initial labels from $\mathcal{P}_{FG}$ and $\mathcal{P}_{BG}$ are propagated to them via random walk [45] with a transition probability matrix, *i.e.*, a matrix of pairwise semantic similarity between points in the input image. In detail, points in $\mathcal{P}_{Box}$ that are highly likely to be propagated from those in $\mathcal{P}_{FG}$ but not from those $\mathcal{P}_{BG}$ are considered as pseudo foreground. Conversely, points in $\mathcal{P}_{Box}$ that are more likely to be propagated from $\mathcal{P}_{BG}$ than $\mathcal{P}_{FG}$ are considered as pseudo background.

3.2.1 Constructing Transition Probability Matrix

To capture the semantic similarity between points, EXITS leverages an attention matrix obtained from a multi-head

self-attention (MHSA) of a ViT encoder. Since the attention matrix of a randomly initialized or ImageNet-pretrained ViT is not capable of discriminating between foreground and background, we warm-up an extra pretrained ViT encoder, called *similarity extractor*, that is additionally trained for only a few epochs on the target dataset with the multiple instance learning (MIL) loss [23, 57]; the loss is defined as

$$
\begin{aligned}
\mathcal{L}_{mil} = \mathcal{L}_{dice}\big(\text{Proj}_x(\mathbf{M}), \text{Proj}_x(\hat{\mathbf{Y}}_{box})\big) \\
+ \mathcal{L}_{dice}\big(\text{Proj}_y(\mathbf{M}), \text{Proj}_y(\hat{\mathbf{Y}}_{box})\big) ,
\end{aligned}
\tag{1}
$$

where $\mathbf{M} \in [0,1]^{H \times W}$ is a mask prediction, $\hat{\mathbf{Y}}_{box} \in \{0,1\}^{H \times W}$ is the area of the bounding box, $\mathcal{L}_{dice}$ indicates the dice loss [53], and $\text{Proj}_x : \mathbb{R}^{H \times W} \mapsto \mathbb{R}^W$ and $\text{Proj}_y : \mathbb{R}^{H \times W} \mapsto \mathbb{R}^H$ are projection operations that apply the max operation across each column and each row vector of the input matrix, respectively. Once trained, the similarity extractor is frozen and used to compute the transition probability matrix during training of the pseudo label generator.

We treat each point as a node in a fully connected graph and construct the transition probability between these nodes using their semantic similarity. To compute the transition probability matrix, a cropped image is divided into $N \times N$ patches and flattened, then fed into the similarity extractor. The similarity matrix $\mathbf{S} \in \mathbb{R}^{N^2 \times N^2}$ is then derived by averaging the self-attention matrices from multiple attention heads of a transformer layer. To construct transition probability matrix $\mathbf{T}$ a doubly stochastic form, the Sinkhorn

Normalization is applied to $\mathbf{S}$, which is calculated by

$$\mathbf{T} = \frac{\mathbf{A} + \mathbf{A}^\top}{2}, \text{ where } \mathbf{A} = \text{Sinkhorn}(\mathbf{S}) , \qquad (2)$$

where $\text{Sinkhorn}(\cdot)$ is the Sinkhorn-Knopp algorithm [30].

Building the transition probability matrix using MHSA offers two advantages. Firstly, since MHSA captures high-level semantic relationship between points, the resulting transition probability matrix prevents points from being propagated to other points with a similar appearance but different semantics. Secondly, MHSA calculates similarities for all point pairs, thereby naturally yielding a transition probability matrix for a fully connected graph. This allows the propagation of labels across separated segments of an object, enhancing the accuracy of the label assignment process.

3.2.2 Generating Pseudo Point Supervision

A pseudo label of $\mathbf{p}_i \in \mathcal{P}_\text{Box}$ is assigned by its propagation score calculated by random walk with the transition probability from each member of $\mathcal{P}_\text{FG}$ and $\mathcal{P}_\text{BG}$ to $\mathbf{p}_i$. We define the foreground propagation score $\pi_i^{(f)}$ of $\mathbf{p}_i$ as

$$\pi_i^{(f)} = \frac{1}{|\mathcal{P}_\text{FG}|} \sum_{\mathbf{p}_j \in \mathcal{P}_\text{FG}} \mathbf{T}^\alpha(j, i) , \qquad (3)$$

where $\mathbf{T}^\alpha(j, i)$ denotes the transition probability that point $\mathbf{p}_j$ propagates to point $\mathbf{p}_i$ through α hops in random walk. The background propagation score of $\mathbf{p}_i$ is defined in an analogous manner,

$$\pi_i^{(b)} = \frac{1}{|\mathcal{P}_\text{BG}|} \sum_{\mathbf{p}_j \in \mathcal{P}_\text{BG}} \mathbf{T}^\alpha(j, i) . \qquad (4)$$

Using these scores, the set of pseudo foreground points $\hat{\mathcal{P}}_\text{FG}$ and that of pseudo background points $\hat{\mathcal{P}}_\text{BG}$ are defined as

$$\hat{\mathcal{P}}_\text{FG} := \left\{ (x_i, y_i) : \exists (x_i, y_i) \in \mathcal{P}_\text{Box}, \pi_i^{(f)} - \pi_i^{(b)} \geq \tau_\text{FG} \right\}$$
$$\hat{\mathcal{P}}_\text{BG} := \left\{ (x_i, y_i) : \exists (x_i, y_i) \in \mathcal{P}_\text{Box}, \pi_i^{(f)} - \pi_i^{(b)} \leq \tau_\text{BG} \right\} , \qquad (5)$$

where τ_FG and τ_BG are threshold hyperparameters.

Point dropout. To enhance the diversity of the pseudo point supervision and prevent overfitting, we adopt an augmentation technique called *point dropout*. For each epoch, point dropout independently eliminates a random subset from both $\hat{\mathcal{P}}_\text{FG}$ and $\hat{\mathcal{P}}_\text{BG}$, and the removed subsets are excluded from the training process during that epoch.

3.2.3 Training Objective

Point loss. Let (x_i, y_i) denote the 2D coordinates of point $\mathbf{p}_i$. We construct sparse binary mask $\hat{\mathbf{Y}} \in \mathbb{R}^{N \times N}$ as follows:

$$\hat{\mathbf{Y}}(x_i, y_i) = \begin{cases} 1 & \text{if } \mathbf{p}_i \in \mathcal{P}_\text{FG} \cup \hat{\mathcal{P}}_\text{FG} \\ 0 & \text{otherwise .} \end{cases} \qquad (6)$$

Furthermore, we construct a masking matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$, which encodes region with point-supervision as follows:

$$\mathbf{K}(x_i, y_i) = \begin{cases} 1 & \text{if } \mathbf{p}_i \in \mathcal{P}_\text{FG} \cup \hat{\mathcal{P}}_\text{FG} \cup \mathcal{P}_\text{BG} \cup \hat{\mathcal{P}}_\text{BG} \\ 0 & \text{otherwise .} \end{cases} \qquad (7)$$

We employ the dice loss between $\hat{\mathbf{Y}}$ and the predicted mask probability $\mathbf{M}$. Prior to computing the loss, $\mathbf{M}$ is downsampled to $\tilde{\mathbf{M}} \in [0,1]^{N \times N}$ to match the size with $\hat{\mathbf{Y}}$. Further, we perform an element-wise multiplication between $\tilde{\mathbf{M}}$ and $\mathbf{K}$ so that the loss signal is applied only to the labeled points. In cases where none of the points is retrieved with the point retrieval algorithm, *i.e.*, $|\hat{\mathcal{P}}_\text{FG} \cup \hat{\mathcal{P}}_\text{BG}| = 0$, we apply the MIL loss in Eq. (1) additionally. The point loss is defined as:

$$\mathcal{L}_\text{point} = \mathcal{L}_\text{dice}(\tilde{\mathbf{M}} \odot \mathbf{K}, \hat{\mathbf{Y}}) + \mathbb{1}_{\{|\hat{\mathcal{P}}_\text{FG} \cup \hat{\mathcal{P}}_\text{BG}| = 0\}} \lambda_\text{mil} \mathcal{L}_\text{mil} , \quad (8)$$

where $\odot$ is harmard-product operator, $\mathbb{1}$ is indicator function, and λ_mil is a balancing hyper-parameter.

Conditional random field loss. To further refine the predicted mask, EXITS employs CRF loss as in [35]. Specifically, EXITS utilizes a teacher network obtained by the exponential moving average of training network, *i.e.*, ViT encoder and mask decoder in pseudo labeled generator parameters. Subsequently, mask predictions from both the training network and the teacher network are averaged to obtain $\mathbf{M}^\text{avg}$. Then, $\mathbf{M}^\text{avg}$ is refined through CRF [33] by using the mean-field algorithm [31] and utilized as pseudo ground-truth mask using the dice loss as follows:

$$\mathcal{L}_\text{crf} = \mathcal{L}_\text{dice}(\mathbf{M}, \text{CRF}(\mathbf{M}^\text{avg})) , \qquad (9)$$

where $\text{CRF}(\cdot)$ is the CRF operation. This approach enables the network to yield a more detailed object mask progressively.

In summary, the overall loss function of EXITS is

$$\mathcal{L}_{overall} = \lambda_\text{point} \mathcal{L}_\text{point} + \lambda_\text{crf} \mathcal{L}_\text{crf} , \qquad (10)$$

where λ_point and λ_crf are balancing hyper-parameters.

3.3. Learning a Fully Supervised Model

In the second stage, EXITS employs the trained pseudo label generator to create pseudo mask labels that serve as ground-truth labels for training a fully supervised instance segmentation model. To generate the pseudo mask labels, images containing k instances are cropped around the corresponding extreme point annotations and fed into the generator, yielding a pseudo mask per object. The decoupled design of the instance segmentation and pseudo labeling models allows for our pseudo labels to be seamlessly integrated into any fully supervised instance segmentation model.

Method	Sup	Backbone	InstSeg Model	Mask AP$_{val}$	Mask AP$_{test}$	(%)Ret.$_{val}$	(%)Ret.$_{test}$
fully-supervised methods							
SOLOv2 [59]	M	ResNet-50	SOLOv2	37.5	38.4	-	-
CondInst [56]	M	ResNet-50	CondInst	-	37.7	-	-
FastInst [19]	M	ResNet-50	FastInst	-	38.6	-	-
SOLOv2 [59]	M	ResNet-101-DCN	SOLOv2	41.7	41.8	-	-
SOLOv2 [59]	M	ResNeXt-101-DCN	SOLOv2	42.4	42.7	-	-
Mask2Former [9]	M	Swin-Small [43]	Mask2Former	46.1	47.0	-	-
weakly-supervised methods							
DiscoBox [34]	B	ResNet-50	SOLOv2	30.7	32.0	81.9	83.3
BoxTeacher [13]	B	ResNet-50	CondInst	-	35.0	-	92.8
MAL [35]	B	ResNet-50	SOLOv2	35.0	35.7	93.3	93.0
EXITS (Ours)	E	ResNet-50	SOLOv2	**36.1**	**36.9**	**96.3**	**96.1**
BoxInst [57]	B	ResNet-101-DCN	CondInst	-	35.0	-	-
DiscoBox [34]	B	ResNet-101-DCN	SOLOv2	35.3	35.8	84.7	85.9
BoxLevelSet [39]	B	ResNet-101-DCN	SOLOv2	35.0	35.4	83.9	83.5
BoxTeacher [13]	B	ResNet-101-DCN	CondInst	-	37.6	-	-
SIM [38]	B	ResNet-101-DCN	CondInst	-	37.4	-	-
MAL [35]	B	ResNet-101-DCN	SOLOv2	38.2	38.7	91.6	92.6
EXITS (Ours)	E	ResNet-101-DCN	SOLOv2	**39.8**	**40.2**	**95.4**	**96.2**
DiscoBox [34]	B	ResNeXt-101-DCN	SOLOv2	37.3	37.9	88.0	88.8
MAL [35]	B	ResNeXt-101-DCN	SOLOv2	38.9	39.1	91.7	91.6
EXITS (Ours)	E	ResNeXt-101-DCN	SOLOv2	**40.5**	**40.9**	**95.5**	**95.8**
MAL [35]	B	Swin-Small [43]	Mask2Former	43.3	44.1	93.9	93.8
EXITS (Ours)	E	Swin-Small [43]	Mask2Former	**44.2**	**45.0**	**95.9**	**95.7**

Table 1. Results on COCO `val2017` and `test-dev`. We report performance using Mask Average Precision (Mask AP) and Retention rate (Ret, %). Retention rate is the performance ratio compared to its fully supervised counterpart. Each method is trained with the supervision of either a mask (M), bounding box (B), or extreme points (E). Note that the annotation cost of the bounding box and extreme points are equal.

4. Experiments

4.1. Experimental Setting

Datasets. Our method is evaluated on three instance segmentation datasets: COCO [40], PASCAL VOC [17], and LVIS v1.0 [18]. We utilize the 2017 version of COCO, which contains 115k images for training, 5k for validation, and 20k for testing across 80 classes. For PASCAL VOC, we employ the augmented version that includes 10,582 training and 1,449 validation images across 20 semantic classes. LVIS v1.0 contains 164k images spanning 1200+ categories, and we follow the standard partition for training and validation sets as described in [18]. To obtain extreme point annotations, we follow the protocol described in ExtremeNet [67][1], which converts mask annotations to extreme point annotations.

Evaluation metric. Following previous work [13, 38, 39] we use coco-style Mask AP as an evaluation metric. For COCO and LVIS v1.0 datasets, we additionally report Retention Rate as in MAL [35], which is the ratio of performance compared to its fully supervised counterpart.

Implementation details (first stage). We followed the architecture of MAL [35] for consistent comparison. The Standard ViT-Base [16], pretrained with MAE [22], served as our ViT encoder, paired with an attention-based mask decoder.

---

[1] https://github.com/xingyizhou/ExtremeNet

Method	Backbone	AP	AP$_{50}$	AP$_{75}$
BoxInst [57]	ResNet-50	34.3	59.1	34.2
DiscoBox [34]	ResNet-50	-	59.8	35.5
BoxLevelSet [39]	ResNet-50	36.3	64.2	35.9
SIM [38]	ResNet-50	36.7	65.5	35.6
BoxTeacher [13]	ResNet-50	38.6	66.4	38.7
MAL† [35]	ResNet-50	37.6	64.8	37.9
EXITS (Ours)	ResNet-50	**40.4**	**67.4**	**41.4**
BBTP [23]	ResNet-101	-	58.9	21.6
Arun *et al.* [2]	ResNet-101	-	57.7	31.2
BBAM [36]	ResNet-101	-	63.7	31.8
BoxInst [57]	ResNet-101	36.4	61.4	37.0
DiscoBox [34]	ResNet-101	-	62.2	37.5
BoxLevelSet [39]	ResNet-101	38.3	66.3	38.7
SIM [38]	ResNet-101	38.6	67.1	38.3
BoxTeacher [13]	ResNet-101	40.3	**67.8**	41.3
MAL† [35]	ResNet-101	38.4	65.7	39.1
EXITS (Ours)	ResNet-101	**41.4**	67.7	**42.5**

Table 2. Results on Pascal VOC `val2012`. Symbol "†" denotes the re-implemented results.

The teacher network is derived from the exponential moving average of the model parameters. We employ AdamW optimizer [44] with the learning rate of 1.5×10^{-6}, adjusted

Figure 4. Qualitative comparison of pseudo mask labels on the Separated COCO dataset. (a) Ours, (b) MAL [35], (c) Ground Truth.

by cosine annealing scheduler. An input image is cropped around an object and resized to 512×512, where data augmentation same as MAL is applied. We use MHSA of the 10th transformer layer of the similarity extractor as similarity matrix to construct the TPM. We set the iteration α to 3, the point dropout rate to 0.9, τ_{FG} to 1×10^{-3}, and τ_{BG} to -1×10^{-4}. For COCO and LVIS v1.0 datasets, the similarity extractor is trained for 1 and 10 epochs, respectively. For VOC, the similarity extractor and the pseudo label generator are trained for 8 epochs and 80 epochs, respectively. More details are given in the supplementary materials.

Implementation details (second stage). Various backbone networks and instance segmentation models are adopted for the second stage. For COCO dataset, we employ ResNets [20], ResNeXts [63], Swin Transformer [43] as backbone and SOLOv2 [59] and Mask2Former [9] as instance segmentation model. For VOC dataset, we employ the ResNet backbone and SOLOv2 instance segmentation model. For LVIS v1.0, we employ ResNeXts backbone and Mask R-CNN [21] instance segmentation model. We follow the training configuration of `mmdetection` [7][2].

4.2. Comparisons with State-of-the-art

Results on COCO. In Table 1, we compare the performance of EXITS with the baselines trained with the supervision of either a mask (M), bounding box (B), or extreme points (E), on the COCO dataset. Note that the extreme point has the same labeling cost as the bounding box. EXITS outperforms the box-supervised baselines in every setting across all the compared backbones and instance segmentation models, indicating that EXITS produces high-quality pseudo labels regardless of the backbone or the applied instance segmentation model. Especially with the ResNet-101-DCN backbone, EXITS outperforms the state of the arts such as

---

[2]https://github.com/open-mmlab/mmdetection

Method	Sup	Backbone	Mask AP$_{val}$	$(\%)$Ret.$_{val}$
fully-supervised methods				
Mask R-CNN [21]	M	RNeXt101-32	25.5	-
Mask R-CNN [21]	M	RNeXt101-64	25.8	-
weakly-supervised methods				
MAL [35]	B	RNeXt101-32	23.7	92.9
EXITS (Ours)	E	RNeXt101-32	**24.1**	**94.5**
MAL [35]	B	RNeXt101-64	24.5	95.0
EXITS (Ours)	E	RNeXt101-64	**24.9**	**96.5**

Table 3. Results on LVIS v1.0. Best results are noted as **bold**.

BoxTeacher(+2.6 AP), SIM(+2.8 AP), and MAL(+1.5 AP) by a significant margin. While the baseline method already achieved a retention rate of over 91%, EXITS further narrows the performance gap with its fully-supervised counterparts.

Results on PASCAL VOC. In Table 2, we compare the performance of EXITS with the baselines on the PASCAL VOC dataset. EXITS outperforms the box-supervised baselines with both the ResNet50 and the ResNet101 backbones. Especially with ResNet50 backbone, EXITS shows a significant improvement of 1.8%p, compared to the previous arts. This shows that EXITS predicts higher-quality masks for instance segmentation compared to box-supervised methods.

Results on LVISv1.0. In Table 3, we compare the performance of EXITS with MAL [35] on the LVIS v1.0 dataset. EXTIS clearly outperforms MAL in both AP and Ret, which indicates the effectiveness of utilizing extreme points.

4.3. Pseudo-label Quality Comparison

We evaluate the quality of the generated pseudo mask on COCO and Separated COCO dataset [64] in mIoU. Separated COCO is a subset of COCO and consists of objects whose segmentation masks are separated into multiple parts due to occlusion. In Table 4, we compare the pseudo label quality with MAL [35]. EXITS shows a significant mIoU
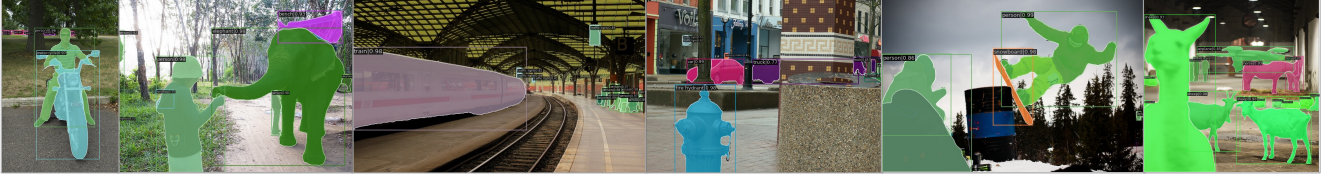
Figure 5. Qualitative results of the final prediction of EXIST on COCO `test-dev`, using Mask2Former with Swin-Small backbone. Our generated pseudo mask labels, EXITS produces high-quality segmentation results, even in separated objects or complex scenes.

	COCO (mIoU)	Separated COCO [64] (mIoU)
MAL [35]	79.1	59.3
EXITS (Ours)	**79.4**	**66.6**

Table 4. Pseudo label quality of the first stage.

$\mathcal{P}_{FG} \cup \mathcal{P}_{BG}$	$\hat{\mathcal{P}}_{FG} \cup \hat{\mathcal{P}}_{BG}$	AP	AP_{50}	AP_{75}
✗	✗	37.6	64.8	37.9
✓	✗	38.0	65.3	38.6
✓	✓	**40.4**	**67.4**	**41.4**

Table 5. Ablation study of the effect of points supervision.

w/ Point dropout	AP	AP_{50}	AP_{75}
✗	39.8	67.1	40.4
✓	**40.4**	**67.4**	**41.4**

Table 6. Effect of the point dropout strategy.

improvement of 7.3%p compared to MAL on the Separated COCO dataset, indicating that EXITS generates high-quality masks for separated objects, thanks to its propagation conducted on the fully connected graphs of all points. This shows that EXITS successfully alleviates the side-effect of the bounding box tightness prior. In Fig. 4, we conduct a qualitative comparison of pseudo mask labels, where EXITS exhibits superior pseudo label quality compared to MAL. Thanks to our high-quality pseudo mask labels, the second stage model produces delicate prediction even in separated objects or complex scenes, as illustrated in Fig. 5.

4.4. Ablation Study

For the ablation studies, we employ ResNet50 backbone with the SOLOv2 model evaluated on the PASCAL VOC dataset using coco-style AP, AP_{50}, AP_{75} metrics. More analysis can be found in the supplement.

Contribution analysis of point set in $\mathcal{L}_{point}$. In Table 5, we evaluate the contributions of the initially labeled point set $\mathcal{P}_{FG} \cup \mathcal{P}_{BG}$, and the pseudo labeled point set $\hat{\mathcal{P}}_{FG} \cup \hat{\mathcal{P}}_{BG}$, when training with $\mathcal{L}_{point}$. We consider MAL [35] as a strong baseline without any point supervision (the first-row of Table 5). The improvement from utilizing $\mathcal{P}_{FG} \cup \mathcal{P}_{BG}$ is marginal, showing that using extreme points naïvely is insufficient to utilize their information for segmentation. Pseudo point supervision from $\hat{\mathcal{P}}_{FG} \cup \hat{\mathcal{P}}_{BG}$ gives significant performance improvement of 2.4%p AP, indicating that our point retrieval algorithm is effective.

Effect of point dropout. In Table 6, we show the effectiveness of our point dropout strategy, which leads to 0.6%p improvement in AP.

Visualizations of pseudo points labels. In Fig. 6, we illustrate the generated pseudo point labels from EXITS. Our pseudo point label accurately captures the object area, effectively excluding the background region even in the occluded areas of the separated objects.
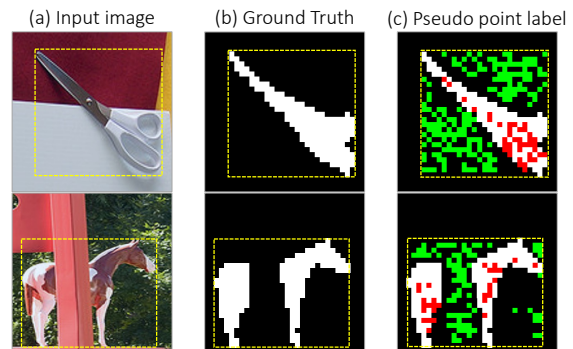

Figure 6. Visualization of pseudo point labels. The white points indicate ground truth, the red indicates $\hat{\mathcal{P}}_{FG}$, and the green points indicates $\hat{\mathcal{P}}_{BG}$. To better visualize pseudo point labels, we use a dropout rate of 0.5 in the illustration. Best viewed in color.

5. Conclusion

We have introduced EXITS, a novel framework for learning instance segmentation using extreme points cost-effectively. EXITS narrows the gap between weakly supervised instance segmentation and its fully supervised counterparts, showing particular strength in segmented objects in severe occlusion scenarios. On the other hand, even with the use of extreme points, differentiating between occluded objects of the same class continues to be a challenging task. Our next agenda is to address this issue by using minimal additional supervision, such as center points.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2209–2218, 2019. 1, 2

[2] Aditya Arun, CV Jawahar, and M Pawan Kumar. Weakly supervised instance segmentation by learning annotation consistent instances. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII*, pages 254–270. Springer, 2020. 6

[3] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019. 2

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: high quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43 (5):1483–1498, 2019. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey 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*, pages 213–229. Springer, 2020. 2

[6] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 7

[8] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 1, 2, 6, 7

[10] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2617–2626, 2022. 1, 3

[11] Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu. Boundary-preserving mask r-cnn. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 660–676. Springer, 2020. 2

[12] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Wenqiang Zhang, Qian Zhang, Chang Huang, Zhaoxiang Zhang, and Wenyu Liu. Sparse instance activation for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4433–4442, 2022. 2

[13] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Qian Zhang, and Wenyu Liu. Boxteacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3145–3154, 2023. 1, 6

[14] Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Solq: Segmenting objects by learning queries. *Advances in Neural Information Processing Systems*, 34: 21898–21909, 2021. 2

[15] Reuben Dorent, Samuel Joutard, Jonathan Shapey, Aaron Kujawa, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Inter extreme points geodesics for end-to-end weakly supervised image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 615–624. Springer, 2021. 3

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. International Conference on Learning Representations (ICLR)*, 2021. 6

[17] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision (IJCV)*, 2010. 2, 6

[18] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 2, 6

[19] Junjie He, Pengyu Li, Yifeng Geng, and Xuansong Xie. Fastinst: A simple query-based model for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23663–23672, 2023. 2, 6

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 7

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 6

[23] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. In *Proc. Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2019. 1, 4, 6

[24] Jie Hu, Chen Chen, Liujuan Cao, Shengchuan Zhang, Annan Shu, Guannan Jiang, and Rongrong Ji. Pseudo-label alignment for semi-supervised instance segmentation. In *Proc.*

IEEE International Conference on Computer Vision (ICCV), pages 16337–16347, 2023. 1

[25] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[26] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Proc. Neural Information Processing Systems (NeurIPS)*, 32, 2019. 1

[27] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 876–885, 2017. 1, 2

[28] Beomyoung Kim, Youngjoon Yoo, Chae Eun Rhee, and Junmo Kim. Beyond semantic to instance segmentation: Weakly-supervised instance segmentation via semantic knowledge transfer and self-refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4278–4287, 2022. 1, 2

[29] Beomyoung Kim, Joonhyun Jeong, Dongyoon Han, and Sung Ju Hwang. The devil is in the points: Weakly semi-supervised instance segmentation via point-guided mask representation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11360–11370, 2023. 1

[30] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008. 5

[31] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011. 5

[32] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)*. 1, 3

[33] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. International Conference on Machine Learning (ICML)*, pages 282–289, 2001. 4, 5

[34] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3406–3416, 2021. 1, 2, 6

[35] Shiyi Lan, Xitong Yang, Zhiding Yu, Zuxuan Wu, Jose M. Alvarez, and Anima Anandkumar. Vision transformers are good mask auto-labelers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23745–23755, 2023. 1, 2, 3, 5, 6, 7, 8

[36] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised

semantic and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2643–2652, 2021. 1, 6

[37] Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, Lionel M Ni, Heung-Yeung Shum, et al. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*, 2022. 2

[38] Ruihuang Li, Chenhang He, Yabin Zhang, Shuai Li, Liyi Chen, and Lei Zhang. Sim: Semantic-aware instance mask generation for box-supervised instance segmentation, 2023. 1, 6

[39] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Xian-Sheng Hua, and Lei Zhang. Box-supervised instance segmentation with level set evolution. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 1–18. Springer, 2022. 1, 2, 6

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. European Conference on Computer Vision (ECCV)*, 2014. 2, 6

[41] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[42] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9819–9828, 2022. 1

[43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6, 7

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[45] László Lovász. Random walks on graphs. *Combinatorics, Paul erdos is eighty*, 2(1-46):4, 1993. 4

[46] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 616–625, 2018. 3

[47] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12674–12684, 2020. 1

[48] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *Proceedings of the IEEE international conference on computer vision*, pages 4930–4939, 2017. 1, 3

[49] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[50] Holger Roth, Ling Zhang, Dong Yang, Fausto Milletari, Ziyue Xu, Xiaosong Wang, and Daguang Xu. Weakly supervised segmentation from extreme points. In *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention: International Workshops, LABELS 2019, HAL-MICCAI 2019, and CuRIOUS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings 4*, pages 42–50. Springer, 2019. 3

[51] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3): 309–314, 2004. 2

[52] Josef Lorenz Rumberger, Jannik Franzen, Peter Hirsch, Jan-Philipp Albrecht, and Dagmar Kainmueller. Actis: Improving data efficiency by leveraging semi-supervised augmentation consistency training for instance segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 3790–3799, 2023. 1

[53] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 4

[54] Chufeng Tang, Lingxi Xie, Gang Zhang, Xiaopeng Zhang, Qi Tian, and Xiaolin Hu. Active pointly-supervised instance segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 606–623. Springer, 2022. 1, 3

[55] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 2

[56] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 282–298. Springer, 2020. 1, 2, 6

[57] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5443–5452, 2021. 1, 2, 4, 6

[58] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 649–665. Springer, 2020. 1, 2

[59] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33: 17721–17732, 2020. 1, 2, 6, 7

[60] Xinggang Wang, Jiapei Feng, Bin Hu, Qi Ding, Longjin Ran, Xiaoxin Chen, and Wenyu Liu. Weakly-supervised instance segmentation via class-agnostic learning with salient images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10225–10235, 2021. 2

[61] Zhenyu Wang, Yali Li, and Shengjin Wang. Noisy boundaries: Lemon or lemonade for semi-supervised instance segmentation? In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16826–16835, 2022. 1

[62] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12193–12202, 2020. 2

[63] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 7

[64] Guanqi Zhan, Weidi Xie, and Andrew Zisserman. A tri-layer plugin to improve occluded detection. *British Machine Vision Conference*, 2022. 2, 7, 8

[65] Gang Zhang, Xin Lu, Jingru Tan, Jianmin Li, Zhaoxiang Zhang, Quanquan Li, and Xiaolin Hu. Refinemask: Towards high-quality instance segmentation with fine-grained features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6861–6869, 2021. 2

[66] Rufeng Zhang, Zhi Tian, Chunhua Shen, Mingyu You, and Youliang Yan. Mask encoding for single shot instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[67] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 850–859, 2019. 3, 6

[68] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3791–3800, 2018. 2

[69] Yanzhao Zhou, Xin Wang, Jianbin Jiao, Trevor Darrell, and Fisher Yu. Learning saliency propagation for semi-supervised instance segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10307–10316, 2020. 1

[70] Yi Zhu, Yanzhao Zhou, Huijuan Xu, Qixiang Ye, David Doermann, and Jianbin Jiao. Learning instance activation maps for weakly supervised instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3116–3125, 2019. 1, 2