Complete Instances Mining for Weakly Supervised Instance Segmentation

Zecheng Li[1] , **Zening Zeng**[1] , **Yuqi Liang**[1] and **Jin-Gang Yu**[1,2*]
[1]South China University of Technology
[2]Pazhou Laboratory
lizecheng19@gmail.com, {zeningzeng, auyqliang}@mail.scut.edu.cn, jingangyu@scut.edu.cn

Abstract

Weakly supervised instance segmentation (WSIS) using only image-level labels is a challenging task due to the difficulty of aligning coarse annotations with the finer task. However, with the advancement of deep neural networks (DNNs), WSIS has garnered significant attention. Following a proposal-based paradigm, we encounter a redundant segmentation problem resulting from a single instance being represented by multiple proposals. For example, we feed a picture of a dog and proposals into the network and expect to output only one proposal containing a dog, but the network outputs multiple proposals. To address this problem, we propose a novel approach for WSIS that focuses on the online refinement of complete instances through the use of MaskIoU heads to predict the integrity scores of proposals and a Complete Instances Mining (CIM) strategy to explicitly model the redundant segmentation problem and generate refined pseudo labels. Our approach allows the network to become aware of multiple instances and complete instances, and we further improve its robustness through the incorporation of an Anti-noise strategy. Empirical evaluations on the PASCAL VOC 2012 and MS COCO datasets demonstrate that our method achieves state-of-the-art performance with a notable margin. Our implementation will be made available at https://github.com/ZechengLi19/CIM.

1 Introduction

Instance segmentation involves the simultaneous estimation of object location and masking segmentation, and it has made significant progress with the assistance of large datasets and instance-level annotations. However, process of obtaining instance-level annotations can be costly and time-consuming. As a solution, weak annotations such as box-level and image-level annotations have been utilized in instance segmentation. Among these weak annotations, the use of image-level annotations is the most cost-effective, but also the most challenging due to the difficulty of aligning coarse annotations with the finer task.

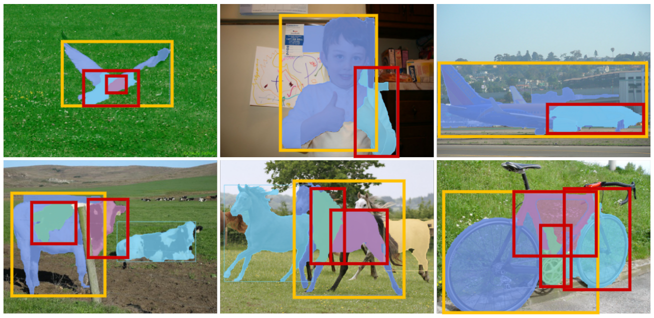

Figure 1: Redundant segmentation. For each instance, it always corresponds to multiple proposals. Yellow boxes: expected segmentations. Red boxes: redundant segmentations.

Weakly supervised instance segmentation (WSIS), which involves using only image-level annotations, has experienced widespread growth in recent years. There are two main paradigms for addressing WSIS: proposal-based paradigm and proposal-free paradigm. The proposal-based paradigm [Zhou *et al.*, 2018; Liu *et al.*, 2020; Shen *et al.*, 2021; Ou *et al.*, 2021] assumes that an instance can be represented by a proposal, thus simplifying WSIS as a classification task. However, these approaches will result in impaired segmentation performance due to redundant segmentation , as illustrated in Figure 1. Typically, if we simply select a few results as output, the segmentation results will be the most discriminative parts of instances. In contrast, the proposal-free paradigm [Ahn *et al.*, 2019; Kim *et al.*, 2022] generates segmentation results online without proposals. During training, this paradigm relies on confident pre-computed pseudo labels to achieve better performance. To leverage these pre-computed pseudo labels, IRNet [Ahn *et al.*, 2019] utilizes Class Attention Maps (CAM) [Zhou *et al.*, 2016] while BESTIE [Kim *et al.*, 2022] uses weakly supervised semantic segmentation (WSSS) maps. The proposal-free paradigm is heavily dependent on pre-computed pseudo labels, which can limit its potential for further improvement.

To address the limitations mentioned above, we propose a novel proposal-based method that operates in an online refinement manner and consists of three key components: MaskIoU heads, a Complete Instances Mining (CIM) strat-

[*]Corresponding Author

egy, and an Anti-noise strategy. Our method also utilizes pre-computed pseudo labels to warm up the model. Instead of using techniques such as random walk (RW) [Ahn *et al.*, 2019] or conditional random fields (CRF) [Ge *et al.*, 2019; Shen *et al.*, 2019; Zhu *et al.*, 2019] to mine complete instances, which can significantly slow down training process, our method employs MaskIoU heads to predict the integrity scores of proposals and the CIM strategy to mine complete instances. We then generate refined pseudo labels based on spatial relationships using the mined complete instances. Finally, the Anti-noise strategy ensures that our method does not suffer from significant performance degradation when pre-computed/refined pseudo labels are noisy.

The main contributions of our work are as follows:

- We introduce the MaskIoU head to WSIS for the first time and propose an Anti-noise strategy to filter out noise caused by pre-computed pseudo labels and refined pseudo labels, improving the robustness of our method.
- We explicitly address the problem called redundant segmentation and present an effective Complete Instances Mining (CIM) strategy to guide the network to pay more attention to complete instances.
- Despite its simplicity, our method achieves state-of-the-art performance on the PASCAL VOC 2012 and MS COCO datasets with a notable margin.

2 Related Work

2.1 Weakly Supervised Object Detection

Multiple Instance Learning (MIL) is a popular paradigm for addressing weakly supervised object detection (WSOD), as it utilizes proposals generated by selective search [Uijlings *et al.*, 2013] as a bag of instances.

WSDDN [Bilen and Vedaldi, 2016] proposed a classification branch and a detection branch to output confident proposals as results, but this approach tends to predict the most discriminative parts of objects. To handle this issue, context-locnet [Kantorov *et al.*, 2016] introduced contextual information to a contrastive model. OICR [Tang *et al.*, 2017] proposed an online instance classifier refinement strategy, which utilizes the proposals of the highest confidence as a pseudo label to supervise the next branch for improved performance. MIST [Ren *et al.*, 2020a] considered the presence of multiple instances and employed Concrete DropBlock to mine complete instances. C-MIL [Wan *et al.*, 2019] designed a novel MIL loss to avoid getting stuck into local minima, while PCL [Tang *et al.*, 2020] transformed the single MIL problem into multiple MIL subproblems through clustering proposals. WSOD^2 [Zeng *et al.*, 2019] combined bottom-up and top-down objectness knowledge as evidence to discover complete instances in the candidates. NDI-WSOD [Wang *et al.*, 2022] and OD-WSCL [Seo *et al.*, 2022] approached WSOD through contrastive learning.

Although WSIS is distinct from WSOD, it is worth noting that our approach was inspired by WSOD approaches and WSOD approaches can be easily transferred to WSIS.

2.2 Weakly Supervised Instance Segmentation

WSIS with only image-level annotations can be divided into two paradigms: proposal-based paradigm and proposal-free paradigm.

Following the proposal-based paradigm, PRM [Zhou *et al.*, 2018] used peaks in Peak Response Maps (PRM) as instance cues and combined them with CAM to predict instances. IAM [Zhu *et al.*, 2019] generated Instance Activation Maps (IAM) through an extent filling module to identify complete instances. Label-PEnet [Ge *et al.*, 2019] employed a cascaded pipeline in a coarse-to-fine manner to simultaneously perform classification, object detection, and instance segmentation. Arun et al. [Arun *et al.*, 2020] used a conditional distribution to explicitly model the uncertainty, which enables pseudo labels to become more reliable. Fan et al. [Fan *et al.*, 2018] and LIID [Liu *et al.*, 2020] used instance saliency labels to support the generation of pseudo labels and considered the relationships in the training set. WS-RCNN [Ou *et al.*, 2021] proposed an Attention-Guided Pseudo Labeling (AGPL) strategy and an Entropic OpenSet Loss to improve WSIS. PDSL [Shen *et al.*, 2021] treated proposals as segmentation cues and proposed a framework for learning detection and segmentation in parallel. Our proposed method is similar to PDSL, but does not require additional segmentation branches due to the ability of proposals to effectively segment instances.

Following the proposal-free paradigm, IRNet [Ahn *et al.*, 2019] used displacement field to indicate instances and generate complete instances via Class Boundary Maps and random walk. BESTIE [Kim *et al.*, 2022] transferred the knowledge of WSSS to WSIS by a strengthening constraint. While our method also leverages pre-computed pseudo labels, we design our approach in an online refinement manner and propose an Anti-noise strategy to reduce the dependence on these pre-computed labels.

Thorough comparisons of representative proposal-based and proposal-free methods show that outputs from these two paradigms focus on different aspects. Proposal-based methods tend to have higher recall and lower precision, while proposal-free methods exhibit the opposite. Despite these differences, there is not a significant gap in performance between the two paradigms. Previous proposal-based methods have effectively utilized the power of proposals, which not only consider bottom-up information of the image but also simplify the task. However, a convolutional neural network (CNN) trained for image classification typically results in redundant segmentation. To alleviate this problem, some researchers [Zhu *et al.*, 2019; Arun *et al.*, 2020; Ou *et al.*, 2021] have made some attempts. In contrast, our method explicitly models this problem and addresses it through the Complete Instances Mining (CIM) strategy.

3 Proposed Method

3.1 Preliminary and Overview

Given an image $\boldsymbol{I}$ sized by $H_I \times W_I$, and its image-level label $Y \in \mathbb{R}^C$, where C is the number of categories and $Y_c = 1$ indicates that image $\boldsymbol{I}$ contains the c^{th} category, and otherwise, $Y_c = 0$. To simplify the task, proposals $R \in \mathbb{R}^{N \times H_I \times W_I}$ are obtained using off-the-shelf proposal techniques [Pont-Tuset

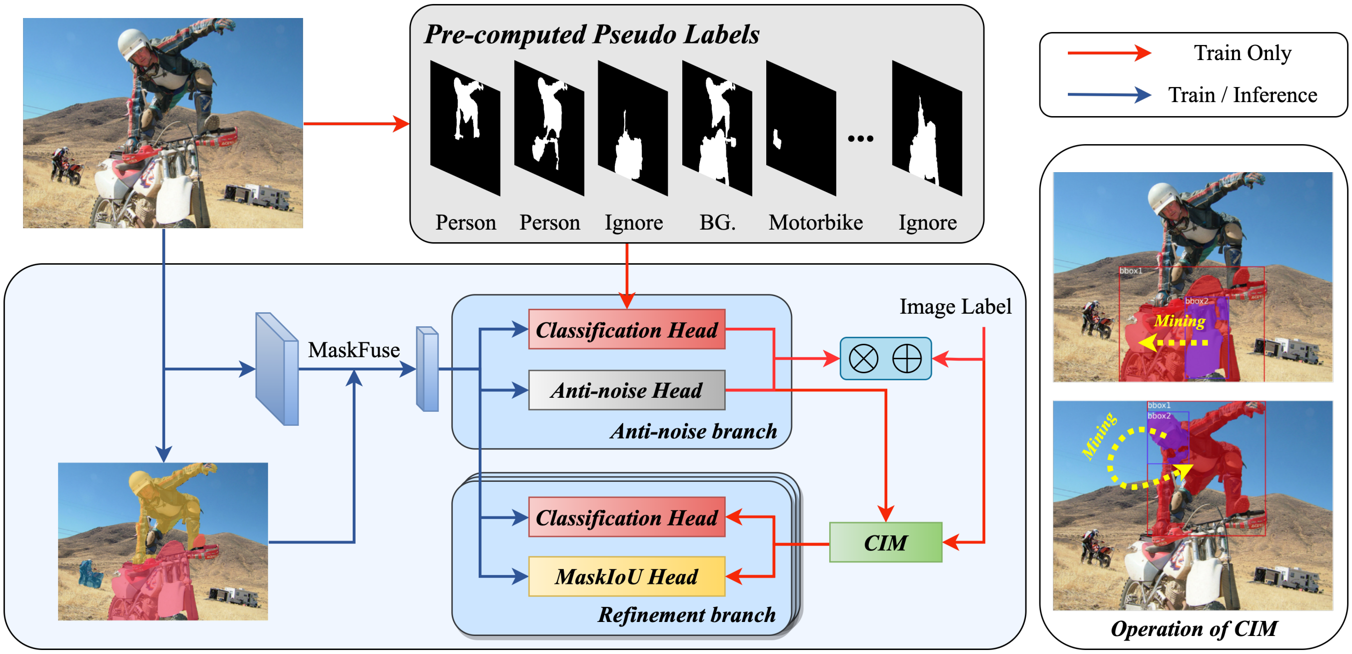

Figure 2: Overview of our proposed method. Our framework mainly contains three components: an Anti-noise branch, K Refinement branches, and Complete Instances Mining (CIM) strategy. Proposal features are generated by MaskFuse and forked into multiple branches. Both Anti-noise and Refinement branches output classification and integrity scores. CIM leverages output of preceding branch to generate refined pseudo labels to supervise next branch, while Anti-noise branch is supervised by pre-computed pseudo labels. In the right column, purple and red represent seeds and pseudo ground truth, respectively. The seeds spread in space to find complete proposals as pseudo ground truth through spatial relationships and integrity scores.

et al., 2017; Maninis *et al.*, 2018], where R_n is a binary segmentation mask and N indicates the number of proposals. In this work, we modify the image-level label $Y \in \mathbb{R}^{C+1}$ to include the background category, denoted as $Y_0 = 1$.

Our method follows the proposal-based paradigm and consists of an Anti-noise branch and several Refinement branches, as illustrated in Figure 2. To begin, we obtain pre-computed pseudo labels using the AGPL strategy proposed by WS-RCNN [Ou *et al.*, 2021]. Then, we use MaskFuse to extract proposal features and feed them into multiple branches. Finally, the CIM strategy explicitly models the redundant segmentation problem and generates refined pseudo labels. To further enhance the robustness of our network, we apply an Anti-noise strategy.

3.2 Reviewing AGPL

The AGPL strategy is a method for generating pseudo labels for a set of proposals in the WS-RCNN model. The method starts by training a classifier and producing CAM. For each target category c, confident peaks from CAM are selected as instance cues, denoted as $p^c = \{p_1^c, p_2^c, \cdots, p_i^c\}$, where p_i^c means the i^{th} peak in the c^{th} category. For each peak, AGPL averages and thresholds proposals containing it to leverage support mask S_i^c, denoted the number of proposals as n_i^c.

$$S_i^c = \left(\frac{1}{n_i^c} \sum_{p_i^c \in R_n} R_n \right) > 0.7, \tag{1}$$

The AGPL strategy sorts the support mask set in descending order according to scores of peaks and assigns labels to proposals based on the IoU between the proposal and the support mask. The pre-computed pseudo labels of proposals are represented as $\hat{y}^0 \in \mathbb{R}^{N \times \{C+1\}}$.

$$\hat{y}_{i,c}^0 = 1 \quad \text{if IoU}(R_i, S_k^c) > 0.5 \tag{2}$$

During the labeling process, each proposal is assigned to only one category. Proposals that overlap with support masks but are not assigned to any categories are assigned as background. Finally, proposal clusters are generated based on whether proposals are assigned by the same support mask, with each background proposal being assigned as a single cluster. The proposal clusters set is denoted as $H = \{(\mathcal{C}_1, M_1, s_1), (\mathcal{C}_2, M_2, s_2), \cdots, (\mathcal{C}_{N^0}, M_{N^0}, s_{N^0})\}$, where $\mathcal{C}_n$, M_n, and s_n represent the set of proposals, the number of proposals, and the label of the n^{th} proposal cluster, respectively.

3.3 MaskFuse

In contrast to WSOD, WSIS utilizes binary masks as proposals rather than bounding boxes. As a result, RoIPool [Girshick, 2015] and RoIAlign [He *et al.*, 2017] operations are not available for WSIS.

To address the lack of feature extraction operation in WSIS, we propose a lightweight operation called MaskFuse. The MaskFuse leverages the bounding box feature $B \in \mathbb{R}^{h \times w \times D}$ obtained through the RoIAlign operation and

Algorithm 1 Complete Instances Mining (CIM) strategy

Input: Image label Y, proposals R, classification scores y^{k-1}, integrity scores t^{k-1}
Parameter: NMS threshold τ_{nms}, containment threshold τ_{con}, percentage of seeds p_{seed}
Output: Pseudo ground truth P^k

```
$P^k = \varnothing$
for $c = 1$ to $C$ do
    if $Y_c = 1$ then
        // Step 1: selecting seeds
        $R_{cls} \leftarrow Sort(y^{k-1}_{*,c})$
        $R_{keep} \leftarrow$ keep top $p_{seed}$ percent of $R_{cls}$
        $ind_{seed} = NMS(R_{keep}, \tau_{nms})$
        // Step 2: mining pseudo ground truth
        Calculate matrix $M \in \mathbb{R}^{N\times N}, M_{i,j} = \frac{R_i \cap R_j}{R_j}$
        $M_{con} = M[:, ind_{seed}] > \tau_{con}$
        $ind_{gt} = argmax(M_{con} t^{k-1}_{*,c}, dim = 0)$
        $P^k_c.append(R[ind_{gt}])$
    end if
end for
```

extracts the corresponding proposal $R_{crop} \in \mathbb{R}^{h\times w}$ using the RoICrop operation. Then, we concatenate $R_{crop} \odot B$ with B and utilize a convolutional layer and two fully connected layers to fuse features from mask-level and box-level. The MaskFuse allows each proposal to be represented by a feature with contextual information, enabling WSOD methods to be transferred to WSIS.

3.4 Refinement Branch

The MaskIoU head, proposed by MS R-CNN [Huang *et al.*, 2019], and the center-ness head, proposed by FCOS [Tian *et al.*, 2019], are designed to learn the quality of predicted results (i.e., masks and bounding boxes) in their respective methods. However, this powerful head is missing in WSIS. To fill this gap, we implement the MaskIoU head into our framework.

Given proposal features, the k^{th} Refinement branch produces two matrixes $y^k, t^k \in \mathbb{R}^{N\times\{C+1\}}$ from classification and MaskIoU heads, which represent the classification and integrity scores, respectively. Furthermore, we employ the CIM strategy to produce refined pseudo labels from preceding branch, denoted as $\hat{y}^k, \hat{t}^k$. Following OICR [Tang *et al.*, 2017], we also propose loss weights $w^k \in \mathbb{R}^N$ to mitigate the degradation caused by noisy refined pseudo labels.

$$\begin{aligned}\mathcal{L}^k_{ref} = &- \frac{1}{N_{fg} + N_{bg}} \sum_{n=1}^{N} \sum_{c=0}^{C} w^k_n \, \hat{y}^k_{n,c} \log y^k_{n,c} \\ &+ \frac{1}{N_{fg}} \sum_{n=1}^{N} \sum_{c=1}^{C} w^k_n \, \hat{y}^k_{n,c} \mathcal{L}_{Smooth-L1}(\hat{t}^k_{n,c} - t^k_{n,c})\end{aligned} \tag{3}$$

where N_{fg} and N_{bg} represent number of foreground and background. $\mathcal{L}_{Smooth-L1}$ indicates smooth-L1 loss [Girshick, 2015].

3.5 Complete Instances Mining (CIM)

Motivated by redundant segmentation, we propose a novel Complete Instances Mining (CIM) strategy. CIM can be divided into two steps: selecting seeds and mining pseudo ground truth. The first step is similar to the operation of MIST [Ren *et al.*, 2020a] strategy.

Let's start with a simple modeling of redundant segmentation. As we expected, partial segmentations typically have high classification scores and low integrity scores, while complete segmentations have the opposite. Additionally, partial segmentations are often spatially contained within complete segmentations. Hence, we utilize classification and integrity scores as metrics for CIM, illustrated in Algorithm 1. CIM awares multiple instances through selecting seeds, which maintains a high recall of pseudo ground truth. For each seed, the proposal that contains it and has the highest integrity score is considered as its corresponding pseudo ground truth. Mining pseudo ground truth enables our method to identify complete instances by aggregating multiple seeds.

To collect more pseudo labels, we assign refined pseudo labels to all proposals. For each proposal, if it highly overlaps with pseudo ground truth, we consider its classification target to be the category of pseudo ground truth. τ_{cls} is a hyper-parameter.

$$\hat{y}^k_{i,c} = \begin{cases} 1 & \text{if } \max(\text{IoU}(R_i, P^k_c)) > \tau_{cls} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The integrity target is determined in the same way. τ_{iou} is also a hyper-parameter.

$$\hat{t}^k_{i,c} = \begin{cases} 1 & \text{if } \max(\text{IoU}(R_i, P^k_c)) > \tau_{iou} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

In which, τ_{iou} is typically greater than τ_{cls}. The loss weight for each proposal can be calculated using Equation 6, where c and j indicate the category and index of pseudo ground truth that has the largest overlap with the R_i, respectively.

$$w^k_i = y^{k-1}_{j,c} \; t^{k-1}_{j,c} \tag{6}$$

We follow the rule mentioned in Section 3.2 to assign background labels. To further improve performance, we implement a cascaded threshold τ_{cas} into our framework inspired by Cascade R-CNN [Cai and Vasconcelos, 2018]. This involves modifying τ_{cls} and τ_{iou} in the k^{th} Refinement branch to $\tau_{cls} + (k-1)\,\tau_{cas}$ and $\tau_{iou} + (k-1)\,\tau_{cas}$, respectively.

3.6 Anti-Noise Strategy

High-quality pre-computed pseudo labels facilitate the network in achieving faster convergence and better performance. However, noise in these labels can confuse the network and lead to overfitting. Similarly, noisy refined pseudo labels also degrade segmentation performance. To address these issues, we propose an Anti-noise strategy comprising an Anti-noise branch and an Anti-noise sampling strategy.

Anti-noise branch. We adopt the WSDDN branch [Bilen and Vedaldi, 2016] as the Anti-noise branch. Similar to the Refinement branches, it also produces two matrixes $y^0, t^0 \in \mathbb{R}^{N\times\{C+1\}}$ from classification head and Anti-noise head. Then, image-level score can be obtained by $y^I =$

Method	Backbone	Sup.	mAP_{25}	mAP_{50}	mAP_{70}	mAP_{75}
Mask R-CNN [He *et al.*, 2017]	ResNet-101	$\mathcal{M}$	76.7	67.9	52.5	44.9
PRM [Zhou *et al.*, 2018]	ResNet-50	$\mathcal{I}$	44.3	26.8	-	9.0
IAM [Zhu *et al.*, 2019]	ResNet-50	$\mathcal{I}$	45.9	28.3	-	11.9
MIST* [Ren *et al.*, 2020a]	VGG-16	$\mathcal{I}$	58.5	43.1	23.2	18.3
Label-PEnet [Ge *et al.*, 2019]	VGG-16	$\mathcal{I}$	49.2	30.2	-	12.9
WS-RCNN [Ou *et al.*, 2021]	VGG-16	$\mathcal{I}$	57.2	42.7	-	19.4
PDSL [Shen *et al.*, 2021]	ResNet50-WS	$\mathcal{I}$	59.3	49.6	-	12.7
BESTIE [Kim *et al.*, 2022]	HRNet-W48	$\mathcal{I}$	53.5	41.8	28.3	24.2
Ours	ResNet-50	$\mathcal{I}$	64.9	51.1	32.4	26.1
Ours	VGG-16	$\mathcal{I}$	65.6	50.8	31.0	25.2
Ours	HRNet-W48	$\mathcal{I}$	**68.3**	**52.6**	**33.7**	**28.4**
WISE† [Laradji *et al.*, 2019]	ResNet-50	$\mathcal{I}$	49.2	41.7	-	23.7
IRNet† [Ahn *et al.*, 2019]	ResNet-50	$\mathcal{I}$	-	46.7	23.5	-
LIID† [Liu *et al.*, 2020]	ResNet-50	$\mathcal{I}, \mathcal{S}$	-	48.4	-	24.9
Arun et al.† [Arun *et al.*, 2020]	ResNet-50	$\mathcal{I}$	59.7	50.9	30.2	28.5
WS-RCNN† [Ou *et al.*, 2021]	VGG-16	$\mathcal{I}$	62.2	47.3	-	19.8
BESTIE† [Kim *et al.*, 2022]	HRNet-W48	$\mathcal{I}$	61.2	51.0	31.9	26.6
Ours†	ResNet-50	$\mathcal{I}$	**68.7**	**55.9**	**37.1**	**30.9**

Table 1: Comparison with the state-of-the-art methods on VOC 2012 dataset. $\mathcal{M}$, $\mathcal{S}$, and $\mathcal{I}$ stand for instance-level, instance saliency, and image-level labels, respectively. † indicates training Mask R-CNN for refinement.

Method	Backbone	Sup.	AP	mAP_{50}	mAP_{75}
		COCO val2017			
Mask R-CNN [He *et al.*, 2017]	ResNet-101	$\mathcal{M}$	35.4	57.3	37.5
WS-JDS [Shen *et al.*, 2019]	VGG-16	$\mathcal{I}$	6.1	11.7	5.5
PDSL [Shen *et al.*, 2021]	ResNet18-WS	$\mathcal{I}$	6.3	13.1	5.0
BESTIE† [Kim *et al.*, 2022]	HRNet-W48	$\mathcal{I}$	14.3	28.0	13.2
Ours	ResNet-50	$\mathcal{I}$	11.9	22.8	11.1
Ours†	ResNet-50	$\mathcal{I}$	**17.0**	**29.4**	**17.0**
		COCO test-dev			
Mask R-CNN [He *et al.*, 2017]	ResNet-101	$\mathcal{M}$	35.7	58.0	37.8
Fan et al.† [Fan *et al.*, 2018]	ResNet-101	$\mathcal{I}, \mathcal{S}$	13.7	25.5	13.5
LIID† [Liu *et al.*, 2020]	ResNet-50	$\mathcal{I}, \mathcal{S}$	16.0	27.1	16.5
BESTIE† [Kim *et al.*, 2022]	HRNet-W48	$\mathcal{I}$	14.4	28.0	13.5
Ours	ResNet-50	$\mathcal{I}$	12.0	23.0	11.3
Ours†	ResNet-50	$\mathcal{I}$	**17.2**	**29.7**	**17.3**

Table 2: Comparison with the state-of-the-art methods on COCO dataset. $\mathcal{M}$, $\mathcal{S}$, and $\mathcal{I}$ stand for instance-level, instance saliency, and image-level labels, respectively. † indicates training Mask R-CNN for refinement.

$\sum_{n=1}^{N}(y^0_{n,*} \odot t^0_{n,*})$. We combine Binary Cross-Entropy and PCL [Tang *et al.*, 2020] losses as the objective function.

$$\mathcal{L}_{anti} = -\frac{1}{C+1}\sum_{c=0}^{C}\{Y_c \log y^I_c + (1-Y_c)\log(1-y^I_c)\} - \frac{\alpha}{N_{fg}+N_{bg}}\sum_{i=1}^{N^0}\sum_{c=0}^{C} M_i\, s_{i,c}\log\sum_{n\in\mathcal{C}_i}\frac{y^0_{n,c}}{M_i} \quad (7)$$

Here, the hyper-parameter α is set to 12 by default. Especially, the Anti-noise head is not directly guided, which enables it to filter out noise in pre-computed pseudo labels.

Anti-noise sampling. Refined pseudo labels also contain noise, as constraints of the CIM strategy are relaxed. Our observations suggest that noisy labels tend to have lower loss weights. Based on this insight, we adopt an Anti-noise sampling strategy that treats loss weights as sampling probabilities to sample pseudo ground truth P^k. Specifically, we im-

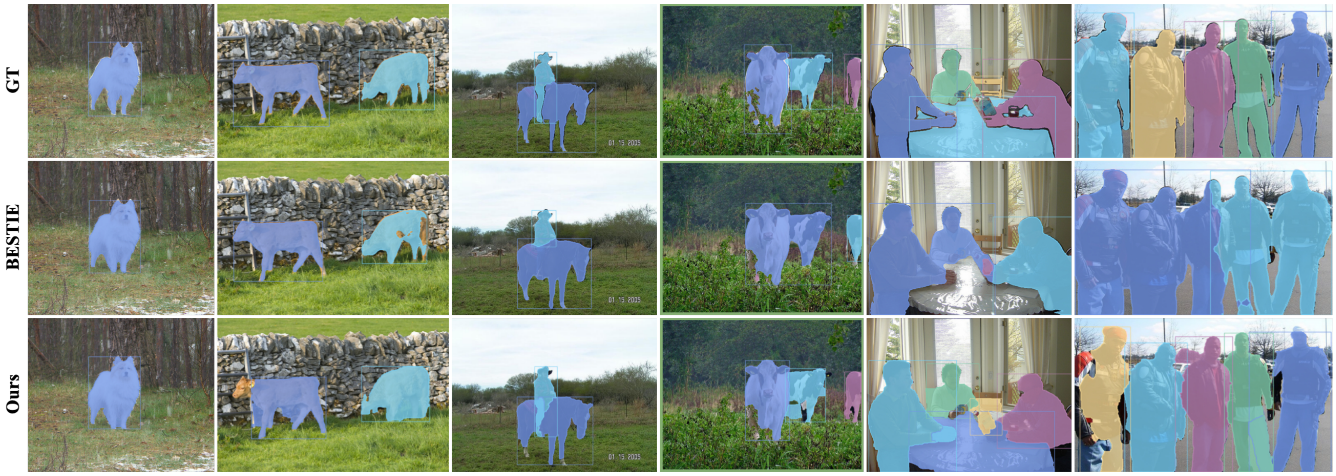

Figure 3: Visualization results on the VOC 2012 dataset. Comparison with BESTIE.

plement sampling with replacement to filter out noise in refined pseudo labels.

In summary, the objective function can be written as:

$$\mathcal{L}_{total} = \mathcal{L}_{anti} + \sum_{k=1}^{K} \mathcal{L}_{ref}^{k} \tag{8}$$

where K is set to 3 by default.

4 Experiments

4.1 Datasets and Evaluation Metrics

Following previous methods, we also evaluate our method on PASCAL VOC 2012 [Everingham *et al.*, 2010] and MS COCO [Lin *et al.*, 2014] datasets. The VOC 2012 dataset includes 10,582 images for training and 1,449 images for validation, comprising 20 object categories. The COCO dataset comprises 115K training, 5K validation, and 20K testing images across 80 object categories. Following previous methods, we report the mean average precision (mAP) with IoU thresholds of 0.25, 0.5, 0.7, and 0.75 for VOC 2012 and report mAP with IoU thresholds from 0.5 to 0.95 for COCO.

4.2 Implementation Details

Our method is implemented in PyTorch and experiments are conducted on an Nvidia RTX 3090. We use COB [Maninis *et al.*, 2018] method to generate proposals for all experiments and utilize ResNet50 [He *et al.*, 2016] as the backbone. As we using mAP_{25} and mAP_{50} as evaluation metrics, we set classification τ_{cls} and integrity τ_{iou} thresholds to 0.25 and 0.5, respectively. The cascaded threshold τ_{cas} is set to 0.1. τ_{nms}, and p_{seed} in Algorithm 1 are set to τ_{cls}, and 0.1, respectively. Containment threshold τ_{con} is set to 0.85 following SoS [Sui *et al.*, 2021]. Although there are many hyper-parameters, we only tune values of τ_{cas} and p_{seed}.

During training, we use the SGD optimization algorithm with an initial learning rate of 2.5×10^{-4} and a weight decay of 5×10^{-4}. We adopt a step learning rate decay schema with a decay weight of 0.1 and set the mini-batch size to 4. The total number of training iterations is 4.5×10^{4} for the VOC 2012 dataset and 24×10^{4} iterations for the COCO dataset. For data augmentation, we apply five image scales $\{480, 576, 688, 864, 1200\}$ with random horizontal flips for both training and testing. During testing, we employ the product of classification and integrity scores as the output of each Refinement branch and average these outputs as the final scores. Following previous methods, we also generate pseudo labels from our method for training Mask R-CNN.

4.3 Comparison With State-of-the-Art

We compare the performance of our method with previous state-of-the-art WSIS methods, as shown in Table 1 and Table 2. For a fair comparison, we also report the results obtained with different backbones, i.e., VGG-16 [Simonyan and Zisserman, 2015] and HRNet-W48 [Wang *et al.*, 2021]. Note that MIST is originally an object detection method, and we adapt it to the WSIS task for comparison.

Our proposed method outperforms all previous methods on both datasets, achieving higher performance than LIID [Liu *et al.*, 2020] even without the use of instance saliency labels, demonstrating that such labels are not necessary for further advancement in WSIS. Additionally, our method demonstrates a 4.9% and 1.7% improvement in terms of mAP_{50} on the VOC 2012 and COCO, respectively, when compared with BESTIE [Kim *et al.*, 2022]. Thanks to impressive semantic segmentation maps produced by WSSS, BESTIE achieves excellent performance on simple images, i.e., single instance, multiple non-adjacent instances, and multiple instances of different categories. However, BESTIE tends to produce grouping instances results caused by its strengthening constraints, illustrated in the second row of Figure 3. In contrast, we follow the proposal-based paradigm which always results in redundant segmentation rather than grouping instances. Hence, we propose CIM to mine complete instances, illustrated in the third row of Figure 3.

Although my approach may appear overly complex, it is implemented using only one linear layer per head, and its training can be completed in a few hours, as shown in Table 3.

Dataset	Backbone	Time (h)
VOC 2012	VGG-16	4.7
VOC 2012	ResNet-50	6.5
VOC 2012	HRNet-W48	30.8
COCO	ResNet-50	37.6

Table 3: Time spent on different configurations.

AGPL	MaskIoU	CIM	Cas	mAP_{50}	mAP_{75}
				38.1	17.5
✔				49.2	20.6
✔	✔			48.8	21.4
✔	✔	✔		50.1	23.8
✔	✔	✔	✔	51.1	26.1

Table 4: Impact of each component: AGPL, MaskIoU head, CIM strategy, and cascaded threshold.

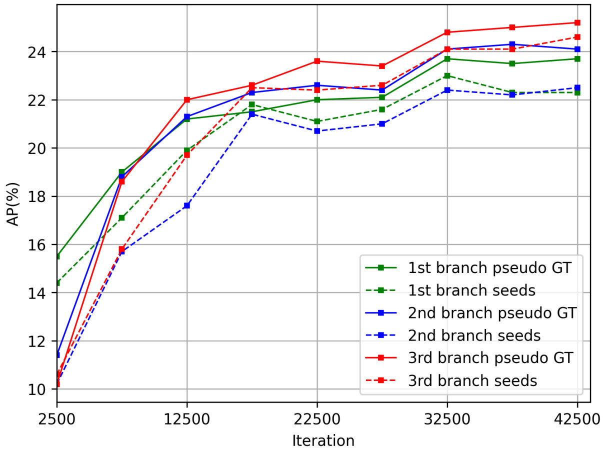

Figure 4: Performance of pseudo GT and seeds generated by CIM.

Meanwhile, the code implementation of CIM is simple.

4.4 Ablation Study

We conduct several ablation studies on the VOC 2012 dataset to evaluate the effectiveness of each component. In these studies, we use ResNet-50 as the backbone and do not apply Mask R-CNN to save time. When the CIM strategy is unavailable, we adopt MIST [Ren *et al.*, 2020a] strategy to replace it and adapt τ_{cls} and τ_{nms} to 0.5.

Impact of AGPL

As shown in Table 4, the use of AGPL results in an improvement in segmentation performance, with an increase of 49.2% and 20.6% in terms of mAP_{50} and mAP_{75}, respectively. Although AGPL produces reliable pre-computed pseudo labels, it only provides coarse complete instance cues, leading to a significant improvement in mAP_{50} but little gain in mAP_{75}.

Impact of MaskIoU and CIM

As shown in Table 4, the performance is slightly improved when the MaskIoU heads are applied alone because the distribution of refined pseudo labels does not change. By employing the CIM strategy, our method achieves a 1.3% and 2.4% improvement in terms of mAP_{50} and mAP_{75}, respectively. This demonstrates that our method can effectively mine complete instances, resulting in improved performance for stricter metrics, i.e., mAP_{75}.

τ_{cas}	0.00	0.05	0.10	0.15	0.20
mAP_{50}	50.1	51.6	51.1	50.4	50.5
mAP_{75}	23.8	25.9	26.1	26.8	26.4

Table 5: Impact of the cascaded threshold value.

p_{seed}	0.05	0.10	0.15	0.20
✘	50.5	49.4	48.6	47.8
✔	51.5	51.1	50.9	51.1

Table 6: Impact of Anti-noise sampling. ✘ and ✔ mean without and with sampling, respectively.

Impact of Cascaded Threshold

Furthermore, the cascaded threshold design allows our method to operate in a coarse-to-fine manner. As shown in the last row of Table 4, the cascaded threshold results in a 1.0% and 2.3% improvement in terms of mAP_{50} and mAP_{75}, respectively, as it guides deeper Refinement branches to focus on more complete instances.

We also evaluate the seeds and pseudo ground truth generated by CIM, as shown in Figure 4. It is obvious that pseudo ground truth outperforms seeds and deeper Refinement branch performs better. As reported in the Table 5, the effect of different cascade thresholds is small.

Impact of Anti-Noise Strategy

Following UFO^2[Ren *et al.*, 2020b], we treat the pre-computed pseudo labels as ground truth and apply a KL-divergence loss on the Anti-noise head, which reduces its performance to 50.4% in terms of mAP_{50}. This result suggests that redundant parameters in the Anti-noise head effectively filter out noise in pre-computed pseudo labels.

To further evaluate the impact of Anti-noise sampling, we introduce noise by increasing the number of seeds in the CIM process. As shown in Table 6, the use of Anti-noise sampling increases the robustness of our method.

5 Conclusion

In this paper, we propose a proposal-based approach in an online refinement manner to address the redundant segmentation problem. Our method incorporates the MaskIoU head and utilizes the CIM strategy to mine complete instances without resorting to RW or CRF. Additionally, we propose the Anti-noise strategy to filter out noise in pseudo labels. Our approach demonstrates state-of-the-art performance on the VOC 2012 and COCO datasets. Moving forward, we expect to investigate methods to enhance the robustness of the model without resorting to Anti-noise sampling strategy that may complicate the analysis.

Acknowledgments

This work was funded by Natural Science Foundation of China under Grants No. 62076099 and No. 61703166.

References

[Ahn *et al.*, 2019] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2204–2213, 2019.

[Arun *et al.*, 2020] Aditya Arun, C.V. Jawahar, and M. Pawan Kumar. Weakly supervised instance segmentation by learning annotation consistent instances. In *Proceedings of the European Conference on Computer Vision*, pages 254–270, 2020.

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

[Cai and Vasconcelos, 2018] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018.

[Everingham *et al.*, 2010] 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*, 88(2):303–338, 2010.

[Fan *et al.*, 2018] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Gang Yu, Ralph R Martin, and Shi-Min Hu. Associating inter-image salient instances for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 367–383, 2018.

[Ge *et al.*, 2019] Weifeng Ge, Weilin Huang, Sheng Guo, and Matthew Scott. Label-penet: Sequential label propagation and enhancement networks for weakly supervised instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3344–3353, 2019.

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

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[Huang *et al.*, 2019] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6409–6418, 2019.

[Kantorov *et al.*, 2016] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *Proceedings of the European Conference on Computer Vision*, pages 350–365, 2016.

[Kim *et al.*, 2022] 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 Conference on Computer Vision and Pattern Recognition*, pages 4268–4277, 2022.

[Laradji *et al.*, 2019] Issam H Laradji, David Vazquez, and Mark Schmidt. Where are the masks: Instance segmentation with image-level supervision. In *Proceedings of the British Machine Vision Conference*, 2019.

[Lin *et al.*, 2014] 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 *Proceedings of the European Conference on Computer Vision*, pages 740–755, 2014.

[Liu *et al.*, 2020] Yun Liu, Yu-Huan Wu, Peisong Wen, Yujun Shi, Yu Qiu, and Ming-Ming Cheng. Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1415–1428, 2020.

[Maninis *et al.*, 2018] Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool. Convolutional oriented boundaries: From image segmentation to high-level tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):819–833, 2018.

[Ou *et al.*, 2021] Jia-Rong Ou, Shu-Le Deng, and Jin-Gang Yu. Ws-rcnn: Learning to score proposals for weakly supervised instance segmentation. *Sensors*, 21(10):3475, 2021.

[Pont-Tuset *et al.*, 2017] Jordi Pont-Tuset, Pablo Arbeláez, Jonathan T. Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):128–140, 2017.

[Ren *et al.*, 2020a] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G. Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10595–10604, 2020.

[Ren *et al.*, 2020b] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Alexander G. Schwing, and Jan Kautz. Ufo^2: A unified framework towards omni-supervised object detection. In *Proceedings of the Eu-*

ropean Conference on Computer Vision, pages 288–313, 2020.

[Seo *et al.*, 2022] Jinhwan Seo, Wonho Bae, Danica J Sutherland, Junhyug Noh, and Daijin Kim. Object discovery via contrastive learning for weakly supervised object detection. In *Proceedings of the European Conference on Computer Vision*, pages 312–329, 2022.

[Shen *et al.*, 2019] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 697–707, 2019.

[Shen *et al.*, 2021] Yunhang Shen, Liujuan Cao, Zhiwei Chen, Baochang Zhang, Chi Su, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Parallel detection-and-segmentation learning for weakly supervised instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8198–8208, 2021.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, pages 1–14, 2015.

[Sui *et al.*, 2021] Lin Sui, Chen-Lin Zhang, and Jianxin Wu. Salvage of supervision in weakly supervised object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14207–14216, 2021.

[Tang *et al.*, 2017] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3059–3067, 2017.

[Tang *et al.*, 2020] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1):176–191, 2020.

[Tian *et al.*, 2019] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9627–9636, 2019.

[Uijlings *et al.*, 2013] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[Wan *et al.*, 2019] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2194–2203, 2019.

[Wang *et al.*, 2021] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2021.

[Wang *et al.*, 2022] Guanchun Wang, Xiangrong Zhang, Zelin Peng, Xu Tang, Huiyu Zhou, and Licheng Jiao. Absolute wrong makes better: Boosting weakly supervised object detection via negative deterministic information. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1378–1384, 2022.

[Zeng *et al.*, 2019] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8291–8299, 2019.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[Zhou *et al.*, 2018] 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.

[Zhu *et al.*, 2019] 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 Conference on Computer Vision and Pattern Recognition*, pages 3111–3120, 2019.