WISH: Weakly Supervised Instance Segmentation using Heterogeneous Labels

Hyeokjun Kweon*
Chung-Ang University
hyeokjunkweon@cau.ac.kr

Kuk-Jin Yoon
KAIST
kjyoon@kaist.ac.kr

Abstract

Instance segmentation traditionally relies on dense pixel-level annotations, making it costly and labor-intensive. To alleviate this burden, weakly supervised instance segmentation utilizes cost-effective weak labels, such as image-level tags, points, and bounding boxes. However, existing approaches typically focus on a single type of weak label, overlooking the cost-efficiency potential of combining multiple types. In this paper, we introduce WISH, a novel heterogeneous framework for weakly supervised instance segmentation that integrates diverse weak label types within a single model. WISH unifies heterogeneous labels by leveraging SAM's prompt latent space through a multi-stage matching strategy, effectively compensating for the lack of spatial information in class tags. Extensive experiments on Pascal VOC and COCO demonstrate that our framework not only surpasses existing homogeneous weak supervision methods but also achieves superior results in heterogeneous settings with equivalent annotation costs.

1. Introduction

Instance segmentation [6, 7, 12, 40] is a fundamental task in computer vision that aims to predict the pixels associated with each semantic instance within an image. With advances in learning-based methods, various instance segmentation paradigms have been extensively studied. However, these methods typically rely on fully supervised learning [6, 7, 12, 40], which requires pixel-level segmentation ground truth (GT) and places a significant burden on the annotation process. This challenge is particularly critical in applications such as autonomous driving and robotics, where large-scale, detailed annotations are essential.

To address this issue, Weakly Supervised Instance Segmentation (WSIS) [1, 2, 8, 15, 16, 24–27, 30, 33, 39, 41, 47, 49] has emerged, reducing the annotation burden by utilizing more cost-effective weak labels. Compared to dense pixel-wise annotations (*i.e.*, GT masks), weak labels contain less information but are far easier and cheaper to acquire. Motivated by this cost-efficiency and high practicality, throughout the last decade, several forms of weak supervision have been explored in WSIS, including image-level class tags [1, 16, 47, 49], sparsely labeled points [2, 16, 26, 27, 33, 41], and bounding boxes [8, 15, 24, 25, 30, 39].

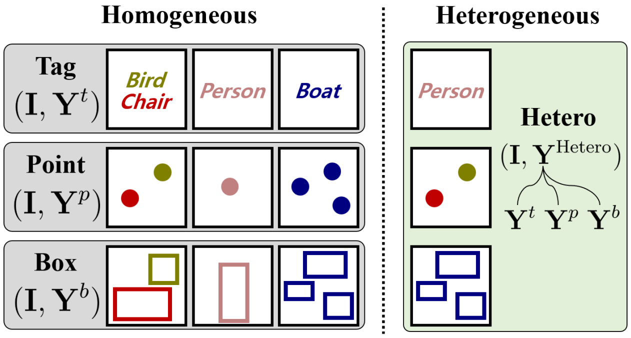

Homogeneous
Heterogeneous
Tag
$(\mathbf{I}, \mathbf{Y}^t)$
Bird
Chair
Person
Boat
Point
$(\mathbf{I}, \mathbf{Y}^p)$
Box
$(\mathbf{I}, \mathbf{Y}^b)$
Person
Hetero
$(\mathbf{I}, \mathbf{Y}^{\text{Hetero}})$
$\mathbf{Y}^t$ $\mathbf{Y}^p$ $\mathbf{Y}^b$

Figure 1. Comparison of homogeneous and heterogeneous settings. Left shows homogeneous setting, where every image in dataset is annotated with a single weak label type. The right shows the proposed heterogeneous WSIS, a generalized approach that allows different weak label types across samples for a single dataset.

Despite these efforts, most existing WSIS methods focus on improving segmentation performance using a single type of weak label, which we term the **homogeneous** setting. In practical scenarios, however, there is no inherent reason to restrict supervision to a single weak label type. Instead, leveraging multiple types of weak labels in a **heterogeneous** manner, as in Fig. 1, can be more effective, considering their varying costs and complementary information. For example, while class tags are the most accessible, they provide no localization information, whereas bounding boxes, though more costly, offer much richer localization cues. Therefore, a strategy that labels part of the training data with class tags and other parts with bounding boxes could be a more cost-effective solution. Yet, the community lacks studies exploring this heterogeneous setting.

In this context, this paper makes two key contributions to advance the field of WSIS. First, we propose a novel heterogeneous framework, named WISH, that simultaneously leverages multiple types of weak labels, generaliz-

[*] This work has been done while at KAIST.

ing and unifying the conventional homogeneous approaches typically used in weakly supervised learning. Inspired by promptable segmentation in SAM [17], WISH treats weak labels themselves as conditions for instance segmentation directly derived from the image. Specifically, we design a method that maps the weak labels annotated for each image into a unified latent space of weak labels through SAM's prompt encoder, and directly predicts each latent vector during the instance segmentation learning process. Notably, WISH framework outperforms existing WSIS methods in each homogeneous setting (for each label type) on Pascal VOC [10] and COCO [34] benchmarks.

Second, through the WISH framework, we explore optimal annotation strategies in heterogeneous settings. Specifically, we propose a novel protocol to test and compare various combinations of heterogeneous weak labels under a fixed budget. Under this protocol, we identify a heterogeneous combination showing substantially better performance than the conventional homogeneous setting, with equivalent annotation costs. While the optimal ratios and combinations may vary depending on the dataset and application domain, this pioneering approach could offer valuable guidelines for WSIS applications across various fields.

To sum up, we not only propose the WISH framework, which achieves outstanding performance, but also highlight an essential engineering challenge overlooked by prior research in our field. We hope that this work serves as a pioneering step toward more practical and effective WSIS.

2. Related Works

2.1. Weakly Supervised Instance Segmentation

Weakly supervised instance segmentation (WSIS) [1, 2, 8, 15, 16, 24–27, 30, 33, 39, 41, 47, 49] aims to alleviate the intensive annotation requirements of fully supervised approaches. As alternatives to pixel-wise mask annotations, prior research has explored various types of weak labels, including image-level classification labels (class tags), points within each instance, and bounding boxes. Specifically, WSIS methods utilizing class tags [1, 16, 47, 49] compensate for the lack of explicit spatial information by employing class activation maps (CAMs) [46] to localize regions corresponding to each class within an image and further separate individual instances. In contrast, methods based on point weak labels [2, 9, 16, 26, 27, 33, 41] leverage precise localization cues from point supervision to identify instance boundaries and use mechanisms such as instance cue identification and self-correction [16] to expand semantic information effectively. Finally, box-based approaches [8, 15, 24, 25, 30, 39] utilize the relatively rich information contained in bounding boxes, applying techniques such as multiple-instance learning [24, 30] and autolabeling [25].

2.2. Use of Multiple Types of Weak Labels

The proposed heterogeneous setting assumes that each image in the training dataset may be annotated with a different type of weak label, resulting in a more general and unified formulation than the conventional homogeneous setting. While few previous studies have considered settings similar to ours, key differences remain. First, there are studies [2, 3, 13, 35, 37] where part of the dataset is annotated strongly (with pixel-wise GT masks), while the rest is annotated weakly. However, these studies are closer to traditional semi-supervised setting with additional weak labels, rather than exploring heterogeneous weak labels. In fact, they typically support only a single type of weak label at a time. Some research has also addressed multiple homogeneous settings simultaneously [14, 16]. However, these approaches generally require specialized modules and hyperparameters for each setting, which limits their applicability in a truly unified heterogeneous setting. In the medical field, where practical approaches are particularly necessary, a method has been proposed that uses both cell-type and position labels for segmentation [36]. However, this approach assumes that each training image includes both types of annotations, unlike our setting, where each image may have only one of several possible types of weak labels. Finally, there is a study about saliency detection using categories and captions [45], which can be considered as heterogeneous. However, the weak labels in this work provide primarily semantic information and are quite similar to each other, limiting their extensibility compared to the diverse weak label types traditionally explored in WSIS.

2.3. SAM in Weakly Supervised Segmentation

SAM [17] is a vision foundation model designed for segmentation tasks. Fascinated by its high performance and generalizability, recent efforts have increasingly focused on leveraging SAM in weakly supervised segmentation. One line of research utilizes SAM for direct inference [4, 5, 11, 38, 48], where SAM is used to improve results obtained from traditional weakly supervised segmentation methods or to predict masks by conditioning SAM on localizations generated from detection models. However, this approach faces limitations in resolving the inherent ambiguity in segmentation tasks and struggles to overcome noise present in the initial predictions [20]. In response, another line of research incorporates SAM into the training process [20, 23, 41, 42]. These studies transfer useful segmentation knowledge from SAM [20] or integrate parts of SAM's model into the training pipeline to adapt other modules accordingly [41, 42]. Although the proposed WISH framework shares a similar philosophy with these approaches, it distinguishes itself by harnessing SAM to simultaneously utilize heterogeneous weak labels for WSIS, unifying them within a single universal framework.

3. Preliminaries

3.1. Problem Definition

We begin by establishing the notations and formally defining the problem targeted in this work. This includes an overview of the conventional problem definition in WSIS studies and a clarification of our specific setting.

Following the protocol in conventional instance segmentation studies, we define the training dataset as

$$\mathbf{D} = \{(\mathbf{I}_1, \mathbf{Y}_1), \dots, (\mathbf{I}_N, \mathbf{Y}_N)\}, \tag{1}$$

where $\mathbf{I}_i \in [0,1]^{3 \times H \times W}$ represents the typical RGB image data and $\mathbf{Y}_i$ is the corresponding instance label. And N is the total number of training data.

In a fully supervised approach [6, 7, 12, 40], each $\mathbf{Y}_i$ provides pixel-wise annotations of instances in $\mathbf{I}_i$, represented as a set of segmentation masks:

$$\mathbf{Y}_i = \{(\mathbf{M}_i^1, \mathbf{c}_i^1), \dots, (\mathbf{M}_i^{k_i}, \mathbf{c}_i^{k_i})\}, \tag{2}$$

where k_i denotes the number of instances in $\mathbf{I}_i$. Each instance is annotated with a binary mask $\mathbf{M}_i^j \in \{0,1\}^{H \times W}$ and a class label $\mathbf{c}^j \in \mathbb{C}$, where $\mathbb{C} = \{1, \dots, C\}$ represents the set of instance segmentation classes.

However, as mentioned in Sec. 1, such pixel-level annotations are costly, posing a significant bottleneck in practical application of instance segmentation methods. WSIS aims to relieve this, training an instance segmentation model using only weak labels, which are more affordable and accessible. In this setting, each training image is annotated with weak labels—such as class tags, points, or bounding boxes—instead of pixel-wise masks.

For WSIS using tags [1, 16, 47, 49], $\mathbf{Y}_i^t$ is defined as

$$\mathbf{Y}_i^t = \{c \mid c \in \mathbb{C}, \text{instance of class } c \text{ exists in } \mathbf{I}_i\}, \tag{3}$$

where t indicates a class tag. Unlike the full annotations in Equ. 2, class tags lack information about the number of instances per class, introducing additional challenges.

Meanwhile, for point annotations [16, 26, 27, 33, 41], $\mathbf{Y}_i^p$ is represented as

$$\mathbf{Y}_i^p = \{(\mathbf{X}_i^1, \mathbf{c}_i^1), \dots, (\mathbf{X}_i^{k_i}, \mathbf{c}_i^{k_i})\}, \tag{4}$$

where $\mathbf{X}_i^j = (x_i^j, y_i^j)$ denotes the pixel coordinates of the annotation point for the j-th instance. Each point lies within its corresponding instance, so we assume that $\mathbf{M}_i^j(x_i^j, y_i^j) = 1$.

Finally, for box annotations [8, 15, 24, 25, 30, 39], $\mathbf{Y}_i^b$ is represented as

$$\mathbf{Y}_i^b = \{(\mathbf{B}_i^1, \mathbf{c}_i^1), \dots, (\mathbf{B}_i^{k_i}, \mathbf{c}_i^{k_i})\}, \tag{5}$$

where $\mathbf{B}_i^j = (x_i^j, y_i^j, h_i^j, w_i^j)$ specifies the bounding box in the "left, top, width, height" format for the j-th instance.

While various types of weak labels have been explored in this field, most existing works rely on a single type of weak label. We refer to this conventional setting as **homogeneous WSIS**, in which each $\mathbf{Y}_i$ is homogeneous.

In contrast, this work introduces the use of *multiple types of weak labels simultaneously*, a generalized form of the homogeneous setting that offers greater flexibility and practicality. Further, as we will demonstrate, this heterogeneous approach can also lead to improved effectiveness. Here, each training sample is annotated with one of the aforementioned weak label types—class tags, points, or bounding boxes, as shown in Fig. 1. Thus, $\mathbf{Y}_i$ in our approach is heterogeneous, a setting we term **heterogeneous WSIS**:

$$\mathbf{Y}_i^{\text{Hetero}} = \mathbf{Y}_i^t \text{ or } \mathbf{Y}_i^p \text{ or } \mathbf{Y}_i^b. \tag{6}$$

Unless stated otherwise, the term "weak labels" in this paper refers to this heterogeneous setting.

3.2. Segment Anything Model (SAM)

SAM [17] is a foundational vision model designed for segmentation tasks. To mitigate the inherent ambiguity in segmentation (*e.g.*, determining what and how much to segment), SAM introduces promptable segmentation, predicting a mask conditioned on input prompts.

Specifically, given an image $\mathbf{I}$ and prompt $\mathbf{P}$, SAM employs an image encoder $\mathcal{E}_{\text{SAM}}^{\text{img}}$ and a prompt encoder $\mathcal{E}_{\text{SAM}}^{\text{prompt}}$ to extract image features and prompt features, respectively. Here, the image features provide contextual information about the scene, while the prompt features serve as conditions for predicting the mask of a specific region (or object) within the image. This mask prediction is performed using $\mathcal{D}_{\text{SAM}}$, an attention-based mask decoder. This process can be formulated as

$$\mathbf{M}_{\text{SAM}} = \text{SAM}(\mathbf{I}; \mathbf{P}) = \mathcal{D}_{\text{SAM}}(\mathcal{E}_{\text{SAM}}^{\text{img}}(\mathbf{I}), \mathcal{E}_{\text{SAM}}^{\text{prompt}}(\mathbf{P})), \tag{7}$$

where $\mathbf{M}_{\text{SAM}}$ denotes a predicted binary mask corresponding to the given prompt $\mathbf{P}$.

SAM officially supports point and box prompts as inputs, focusing on spatial localization, rather than semantic emphasis as is common in traditional approaches. Interestingly, the input prompts supported by SAM resemble the types of weak labels commonly studied in WSIS. We believe this resemblance is not coincidental; rather, it highlights the suitability and efficiency of such formats as conditioning cues for segmentation targets in an image (usually instances). This natural convergence suggests that prompts themselves may possess substantial expressiveness for representing—and potentially learning—instance segmentation directly from images. In this context, we propose a novel framework that strategically leverages SAM's design focus on prompt to address heterogeneous WSIS, as detailed in the following section.

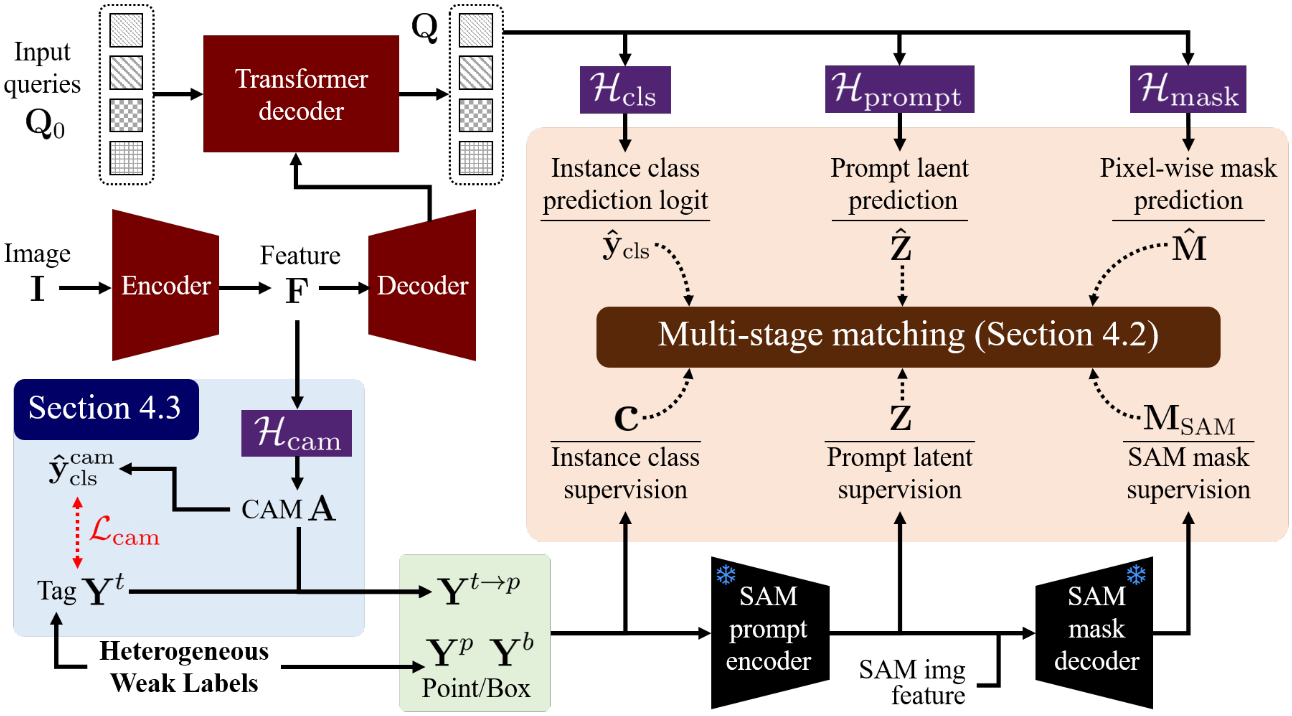

Input
queries
$\mathbf{Q}_0$
Transformer decoder
$\mathbf{Q}$
$\mathcal{H}_{\text{cls}}$
$\mathcal{H}_{\text{prompt}}$
$\mathcal{H}_{\text{mask}}$
Instance class prediction logit
Prompt laent prediction
Pixel-wise mask prediction
$\hat{\mathbf{y}}_{\text{cls}}$
$\hat{\mathbf{Z}}$
$\hat{\mathbf{M}}$
Image
$\mathbf{I}$
Encoder
Feature
$\mathbf{F}$
Decoder
Multi-stage matching (Section 4.2)
$\mathbf{C}$
$\mathbf{Z}$
$\mathbf{M}_{\text{SAM}}$
Instance class supervision
Prompt latent supervision
SAM mask supervision
Section 4.3
$\mathcal{H}_{\text{cam}}$
$\hat{\mathbf{y}}^{\text{cam}}_{\text{cls}}$
CAM $\mathbf{A}$
$\mathcal{L}_{\text{cam}}$
Tag $\mathbf{Y}^t$
$\mathbf{Y}^{t\rightarrow p}$
$\mathbf{Y}^p$ $\mathbf{Y}^b$
Point/Box
Heterogeneous Weak Labels
SAM prompt encoder
SAM img feature
SAM mask decoder

Figure 2. Overview of the proposed **WISH** framework. WISH extends Mask2Former [7] with modules for multi-stage matching, including instance class, prompt latent, and SAM mask supervision (refer to Sec. 4.2). The SAM prompt encoder and mask decoder enable effective mask predictions guided by heterogeneous weak labels. To handle class tags, they are converted to point prompts using CAMs, learned with image-level classification loss $\mathcal{L}_{\text{cam}}$ as described in Sec. 4.3. For better understanding, the dot product with decoder features for pixel-wise mask prediction (refer to Equ. 8) and the self-enhancement loss $\mathcal{L}_{\text{self}}$ have been omitted from this figure.

4. Methods

4.1. WISH Framework

In this paper, we propose a novel framework for **W**eakly supervised **I**nstance **S**egmentation from **H**eterogeneous weak labels, named **WISH**, as visualized in Fig. 2.

Our WISH is designed based on Mask2Former [7]. In WISH, an encoder extracts a feature map $\mathbf{F}$ from the input image. A pixel decoder then derives multi-scale embeddings $\mathbf{S}_{1:L}$ from this feature representation. Meanwhile, the transformer decoder, which plays a central role in mask prediction, provides query features $\mathbf{Q} \in \mathbb{R}^{N_q \times D_q}$ through masked attention between input queries $\mathbf{Q}_0 \in \mathbb{R}^{N_q \times D_q}$ and multi-scale embeddings. The input queries are represented as learnable tensors, where N_q is the number of queries.

In the standard Mask2Former [7], which targets a fully supervised setting, these query features are used to predict both masks and classes. For mask prediction, the mask prediction head $\mathcal{H}_{\text{mask}}$ projects the query features, which are then multiplied with the highest-resolution multi-scale embedding ($\mathbf{S}_L$) to obtain the mask prediction $\hat{\mathbf{M}}$. For class prediction, a classification head $\mathcal{H}_{\text{cls}}$ processes the query features to produce class prediction logit $\hat{\mathbf{y}}_{\text{cls}} \in \mathbb{R}^{C+1}$, including an auxiliary "no object" ($\emptyset$) score as well as the original set of categories ($\mathbb{C}$), following the previous works [6, 7]. These processes are formally denoted as

$$\hat{\mathbf{M}} = \mathcal{H}_{\text{mask}}(\mathbf{Q}) \cdot \mathbf{S}_L, \quad \hat{\mathbf{y}}_{\text{cls}} = \mathcal{H}_{\text{cls}}(\mathbf{Q}). \tag{8}$$

However, unlike fully supervised mask classification, pixel-wise GT is inaccessible in a weakly supervised setting. To address this, we draw inspiration from recent studies [28, 32, 42] that apply constraints in the prompt feature space and introduce an additional prompt prediction head $\mathcal{H}_{\text{prompt}}$. The goal of this head is to leverage the prompt encoder of SAM [17], which is pre-trained to map prompts (such as points or boxes) into the prompt latent space efficiently. Specifically, for the heterogeneous weak label set associated with each input image, we apply the prompt encoder to project each label into the prompt latent space as

$$\mathbf{Z} = \mathcal{E}^{\text{prompt}}_{\text{SAM}}(\mathbf{Y}^{\text{Hetero}}), \tag{9}$$

where $\mathbf{Z} = \{z_1, \ldots, z_k\}$ denotes the set of prompt latent vectors of the input image $\mathbf{I}$, and k is the number of GT instances. Here, the class tag, unlike point or box, is actually incompatible with SAM's prompt encoder. For better understanding, we assume that the class tag can be projected by the prompt encoder somehow for now, and how we address this limitation is discussed in Sec. 4.3.

The prompt prediction head aims to directly predict the GT prompt features $\mathbf{Z}$ from the query feature $\mathbf{Q}$ as

$$\hat{\mathbf{Z}} = \mathcal{H}_{\text{prompt}}(\mathbf{Q}), \tag{10}$$

where $\hat{\mathbf{Z}} = \{\hat{\mathbf{Z}}^1, \ldots, \hat{\mathbf{Z}}^{N_q}\}$. This design strategy effectively transfers the SAM prompt encoder's ability to extract segmentation-relevant conditions into the WISH framework. Note that the number of predicted prompt latent vectors is N_q, which is the size of the input queries set, differs from the number of GT instances (N).

4.2. Multi-Stage Matching for WISH

In a fully supervised setting with Mask2Former [7], the predictions in Equ. 8 are guided by a set of pixel-wise GT masks through bipartite matching. The matching cost comprises a cross-entropy cost for classification and a mask distance. The classification cost between the class prediction of i-th query and the j-th instance (in GT) is defined as

$$\text{Cost}_{\text{cls}}^{i,j} = \hat{\mathbf{y}}_{\text{cls}}^{i}(\mathbf{c}^j), \tag{11}$$

where $\mathbf{c}^j$ is the GT class of j-th instance. Following MaskFormer [6], we pad the set of GT instances with $\emptyset$ tokens.

The proposed WISH framework introduces two key modifications. The first is the use of a prompt matching cost, defined as the Kullback-Leibler Divergence (KLD) between the previously obtained $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ as

$$\text{Cost}_{\text{prompt}}^{i,j} = \text{KLD}(\hat{\mathbf{Z}}^i, \mathbf{Z}^j). \tag{12}$$

This allows the proposed WISH to effectively learn the concept of instance, via matching in the prompt latent space.

The second modification handles the absence of pixel-wise GT masks in the heterogeneous setting. Although the weak labels (*i.e.*, prompts) provide significant segmentation conditioning, they alone are insufficient to achieve the learning of mask predictions at the pixel level.

To address this, we leverage SAM [17] to generate SAM masks corresponding to each weak label. An important trait of SAM is its intentional design to output three candidate masks for each prompt. This choice addresses the inherent ambiguity in segmentation, where the most appropriate mask may vary depending on context. We modify Mask2Former's mask distance function to adaptively select the closest match among the three SAM masks as:

$$\text{Cost}_{\text{mask}}^{i,j} = \min_{n\in\{1,2,3\}} d(\hat{\mathbf{M}}^i, \mathbf{M}_{\text{SAM}}^{j,n}), \tag{13}$$

where $\mathbf{M}_{\text{SAM}}^{i,n}$ denotes the n-th mask of $\text{SAM}(\mathbf{I}; \mathbf{Y}_i)$ and $d(\cdot)$ is the mask distance function used in Mask2Former [7].

To sum up, the total cost for bipartite matching between the predicted instances and the weak supervision is as

$$\text{Cost}^{i,j} = \alpha\text{Cost}_{\text{cls}}^{i,j} + \beta\text{Cost}_{\text{prompt}}^{i,j} + \gamma\text{Cost}_{\text{mask}}^{i,j}. \tag{14}$$

We solve this matching via Hungarian algorithm [18]. Finally, given a matched gt instance for each instance prediction, the segmentation loss for the WISH framework is as

$$\mathcal{L}_{\text{seg}} = \sum_{i=1}^{N_q} \left[-\log \hat{\mathbf{y}}_{\text{cls}}^{i}(\mathbf{c}^{\text{gt}}) + \mathbf{1}_{\mathbf{c}^{\text{gt}}\neq\emptyset} \left(\text{Cost}_{\text{prompt}}^{i,\text{gt}} + \text{Cost}_{\text{mask}}^{i,\text{gt}} \right) \right], \tag{15}$$

where $\mathbf{1}_{\mathbf{c}^{\text{gt}}\neq\emptyset}$ represents that we do not impose prompt or mask loss for the padded GT instances of $\emptyset$ class.

4.3. Handling Class Tags

Unlike other weak labels (*i.e.*, points or boxes), class tags lack explicit spatial information. Consequently, direct localization is not possible with class tags, leading prior weakly supervised works in instance segmentation [1, 16] and semantic segmentation [19, 21, 22, 43, 44] to rely on class activation maps (CAMs) [46]. In line with previous methods [1, 20], we feed the feature map from the image encoder into $\mathcal{H}_{cam}$, a 1×1 convolutional CAM-head, as

$$\mathbf{A} = \mathcal{H}_{\text{cam}}(\mathbf{F}), \tag{16}$$

where $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$ denotes CAMs. Then, we apply Global Average Pooling (GAP) to the CAMs and obtain a logit $\hat{\mathbf{y}}_{\text{cls}}^{\text{cam}} \in \mathbb{R}^{C}$ for image-level classification as $\hat{\mathbf{y}}_{\text{cls}}^{\text{cam}} = \text{GAP}(\mathbf{A})$. Note that $\mathcal{H}_{\text{cam}}$ and $\hat{\mathbf{y}}_{\text{cls}}^{\text{cam}}$ are differ from $\mathcal{H}_{\text{cls}}$ and $\hat{\mathbf{y}}_{\text{cls}}$ in Equ. 8, which are for the query features $\mathbf{Q}$.

Fortunately, all three types of weak labels—class tags, points, and boxes—carry class information about the input image. Therefore, we can directly supervise the image-level classification prediction via typical cross-entropy loss as

$$\mathcal{L}_{\text{cam}} = \mathcal{L}_{\text{ce}}(\hat{\mathbf{y}}_{\text{cls}}^{\text{cam}}, \mathbf{c}^{\text{img}}), \tag{17}$$

where $\mathbf{c}^{\text{img}}$ is the image-level classification GT from the weak label, differs from the class GT of each instance.

The CAMs learned by classification loss can highlight the image regions corresponding to each class, providing a reasonable basis for localization. However, an additional challenge arises: SAM's prompt encoder cannot directly process the continuous score map produced by CAMs. To address this, we propose an approach inspired by recent works using SAM [20, 31], where multiple local peaks are identified in the CAM and used as positive point prompts [47]. Specifically, we apply a local maximum filter on the CAM to obtain a set of local peaks. The peaks below τ are rejected. Details about this process are in *Supp*.

To distinguish instances, we utilize masks generated by using each local peak as a point prompt of SAM. We compute the IoU between these masks, and if the IoU exceeds a given threshold, the corresponding local peaks are considered to belong to the same instance. Through empirical analysis, we found that using multiple local peaks simultaneously or incorporating negative point prompts tended to reduce mask quality. Therefore, for each instance, we selected the local peak with the highest mask stability score.

After these processes, the original class tag weak label in Equ. 3 can be converted into

$$\mathbf{Y}_i^{t\to p} = \{(\mathbf{X}_i^1, \mathbf{c}_i^1), \dots, (\mathbf{X}_i^{k'_i}, \mathbf{c}_i^{k'_i})\}, \tag{18}$$

having the same form with the point weak label in Equ. 4 and k'_i is the number of converted instances. This approach effectively converts the rough localization of CAMs into instance-aware prompts that are compatible with SAM.

Finally, we introduce a self-enhancement loss to dynamically improve the quality of CAMs throughout training. Since CAMs produce an activation map for each class rather than individual instances, we generate a semantic prediction map $\hat{\mathbf{M}}_c$ for class c from instance mask predictions $\hat{\mathbf{M}}$. This is done by merging binary masks $\hat{\mathbf{M}}^j$ such that its the class prediction $\hat{\mathbf{y}}^j_{\text{cls}}$ have the highest score for class c. We use a simple OR operation for merging. The self-enhancement loss is defined as a binary cross-entropy loss between each class's CAM and the corresponding merged mask:

$$\mathcal{L}_{\text{self}} = \mathcal{L}_{\text{bce}}(\mathbf{A}, \hat{\mathbf{M}}_c). \tag{19}$$

To sum up, the total loss function for training the proposed WISH framework is

$$\mathcal{L}_{\text{WISH}} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{self}}. \tag{20}$$

5. Experimental Results

5.1. Settings

5.1.1 Datasets and Metrics

We conduct extensive experiments on two standard benchmarks for WSIS: PASCAL VOC [10] and MS-COCO [34]. For PASCAL VOC, we used the augmented version, consisting of 10,582 training images and 1,449 validation images covering 20 semantic classes. For COCO, we used the 2017 version, which includes 115,000 training images, 5,000 validation images, and 20,000 test images across 80 classes. Following prior studies [8, 16, 27, 33], we use the COCO-style Mask Average Precision (AP) for evaluation.

5.1.2 Implementation Details

To realize the WISH framework, we employ the official implementation[1] of Mask2Former [7], and most hyperparameters also follow it. Notably, we set the weights for class, mask, and prompt in the matching and mask classification loss to 2, 5, and 5, respectively. In Equ. 14, α, β, and γ are set to 2, 5, and 5, respectively. Since the quality of CAMs is poor in the early stages, we initialize WISH by using only $\mathcal{L}_{\text{cam}}$ for the first 3 epochs for PASCAL and 30,000 iters for COCO.

Table 1. Quantitative comparison between WISH and WSIS works under homogeneous settings on VOC 2012 [10] val set. (M,T,P,B) denotes the types of supervision: (Mask, Tag, Point, Box).

Sup	Method	Backbone	**AP**	**AP$_{50}$**	**AP$_{75}$**
M	Mask R-CNN [12]	HRNet	-	67.9	44.9
	Mask R-CNN [12]	R50	-	68.8	43.3
	Mask2Former [7]	R50	54.8	73.2	58.9
T	IRN [1]	R50	-	46.7	23.5
	BESTIE [16]	HRNet	-	51.0	26.6
	WISH	R50	**46.0**	**62.9**	**50.5**
P	BESTIE [16]	HRNet	-	56.1	30.2
	Attnshift [33]	ViT-S	-	57.1	30.4
	SAPNet [41]	R101	-	64.8	58.7
	WISH	R50	**52.4**	**72.3**	**59.2**
B	BoxInst [39]	R50	32.2	58.1	31.0
	BoxInst [39]	R101	34.4	60.1	34.6
	DiscoBox [24]	R50	-	59.8	35.5
	BoxLevelSet [30]	R50	36.3	64.2	35.9
	SIM [29]	R50	36.7	65.5	35.6
	BoxTeacher [8]	R50	38.6	66.4	38.7
	BoxTeacher [8]	R101	40.3	67.8	41.3
	WISH	R50	**54.6**	**76.1**	**59.8**

5.2. Comparisons with Homogeneous SoTAs

The proposed WISH framework, while primarily designed to leverage the heterogeneous setting effectively, also achieves outstanding performance in homogeneous settings, commonly used in conventional WSIS approaches. As shown in Table 1, our method establishes new state-of-the-art (SoTA) results on the PASCAL VOC benchmark across all homogeneous settings (*i.e.*, class tag, point, and box), significantly outperforming the existing methods.

Furthermore, on the more challenging COCO benchmark, Table 2 confirms that WISH still shows its superiority by achieving meaningful performance gains over existing WSIS methods across all homogeneous settings. These results underline the effectiveness of our WISH framework, even under traditional homogeneous WSIS settings.

One key contributor to this performance boost is, unsurprisingly, the integration of SAM, a model known for its powerful segmentation capabilities. Our framework strategically leverages SAM's segmentation strength, providing a clear advantage over previous approaches. We acknowledge that the use of SAM may render direct comparisons somewhat unfair; however, with the recently growing integration of foundation models in computer vision, we believe this is a timely opportunity to explore effective strategies for utilizing SAM, particularly in the context of weakly supervised segmentation. In response to this need, this paper proposes a novel WSIS framework that leverages SAM's segmentation knowledge through multi-stage matching in both prompt latent and mask spaces. While our primary focus is on heterogeneous settings, the ability of our method to handle multiple homogeneous settings within a single unified model highlights its versatility, offering a flexible solution

[1] https://github.com/facebookresearch/Mask2Former

Table 2. Quantitative comparison between WISH and conventional WSIS methods under homogeneous settings on COCO 2017 [34].

Sup	Method	Backbone	mAP$_{\text{val}}$	mAP$_{\text{test}}$
M	SOLOv2 [40]	R50	37.5	38.4
	SOLOv2 [40]	R101-DCN	41.7	41.8
	Mask R-CNN [12]	R50	42.5	-
	Mask2Former [7]	R50	43.0	43.1
T	BESTIE [16]	HRNet	14.3	14.4
	WISH	R50	**20.7**	**20.9**
P	BESTIE [16]	HRNet	17.7	17.8
	Attnshift [33]	HRNet	21.2	21.9
	SAPNet [41]	R50	31.2	-
	WISH	R50	**31.9**	**32.0**
B	BoxInst [39]	R101-DCN	-	35.0
	DiscoBox [24]	R50	30.7	32.0
	DiscoBox [24]	R101-DCN	35.3	35.8
	BoxTeacher [8]	R50	-	35.0
	BoxTeacher [8]	R101-DCN	-	37.6
	BoxLevelSet [30]	R101-DCN	35.0	35.6
	SIM [29]	R101-DCN	-	37.4
	MAL [25]	R50	35.0	35.7
	MAL [25]	R101-DCN	38.2	38.8
	WISH	R50	**42.3**	**42.7**

applicable to a wide range of scenarios. We believe that, with these capabilities, the proposed WISH framework can make a significant contribution to the field.

5.3. Experiments on Heterogeneous Weak Labels

The primary motivation for using weak labels in WSIS is their cost efficiency, as they are weak (*i.e.*, providing less information) yet significantly easier to obtain compared to pixel-wise mask labels. However, how does this efficiency hold when comparing different types of weak labels? As is well known, class tags are less expensive than points, and boxes are the most costly. Conversely, the information they provide follows an opposite trend: class tags offer the least information, while points and boxes provide increasingly better localization. This means there is still a trade-off between cost and performance, even among weak labels. Previous research has largely focused on homogeneous settings, leaving little opportunity to explore this trade-off. In contrast, our heterogeneous setting enables us to address this issue, allowing for a novel and extensive analysis.

Here, we pose an essential engineering question: **Given a fixed budget, what is the optimal labeling strategy for WSIS?** Should it follow a homogeneous setting, or if heterogeneous, what is the ideal proportion of each label type? The answer would depend on factors like annotation costs or dataset characteristics. Nevertheless, the heterogeneous setting proposed in this paper represents a pioneering attempt to address these questions, offering insights for both academic research and practical applications.

To experiment with this heterogeneous setting, we develop a practical protocol that considers the “value of weak labels,” a factor often overlooked in WSIS. Specifically, we define the relative annotation cost for each type of weak label and explore optimal label combinations within a fixed budget ζ. The set of combinations is represented as (N_t, N_p, N_b), a tuple indicating the number of images labeled with tags, points, and boxes. Denoting the annotation costs for each label type as β_t, β_p, and β_b, respectively. Therefore, the combinations permitted in protocol satisfy

$$N_t\beta_t + N_p\beta_p + N_b\beta_b = \zeta. \tag{21}$$

Following this protocol, we conduct quantitative experiments using the WISH framework on the PASCAL VOC 2012 dataset. For reference values of the βs, we use the official pricing of Amazon Mechanical Turk (AMT)[2], one of the most popular annotation services. According to AMT, the relative cost ratio for classification labeling versus bounding box labeling (per instance) is approximately 1:4. Given that each image in the PASCAL dataset contains an average of around three instances, the annotation cost ratio becomes about 1:12. While the exact costs for point annotations are not explicitly listed, box labeling requires two points (top left and bottom right) per object, so we estimated the cost of point annotation as half that of box annotation. Thus, we set the cost ratio as β_t:β_p:β_b = 1:6:12.

For simplicity, the budget in our protocol is defined as the total cost required to annotate all images in the PASCAL VOC [10] train_aug set with class tags. That is, ζ is set to 10582, and some example combinations within this budget are (4582,1000,0) and (1582,500,500). Naturally, this generalized setting includes homogeneous settings as the three extreme cases: (10582,0,0), (0,1764,0), and (0,0,882). Note that, given the fixed budget, the number of images that can be labeled with points or boxes is significantly reduced compared to the original homogeneous setting.

In practice, of course, both the values of βs and the specific combinations permitted in this protocol can vary significantly. For example, in medical segmentation, where semantic identification is particularly challenging, the relative cost of tags may be higher than estimated here. Conversely, for tasks like camouflaged object segmentation, where boundary identification is difficult, box annotation costs may be higher. Thus, rather than focusing on the absolute accuracy of these cost estimates, the value of this protocol lies in its quantitative consideration of cost differences between weak labels in a heterogeneous setting.

Unfortunately, testing all possible combinations would require extensive computational resources. Instead, in this paper, we focus on two primary cases: tag-point and tag-box combinations, as shown in Table 3 (also refer to Supp.).

[2] https://aws.amazon.com/sagemaker-ai/groundtruth/pricing/

Table 3. Results of WISH framework under heterogeneous settings on PASCAL VOC val set. The top row provides the results of tag-point while the bottom row shows the results of tag-box.

N_t	10,582	7,936	5,290	2,644	0
N_p	0	441	882	1,323	1,764
AP	46.0	46.9	**47.3**	47.2	44.7

N_t	10,582	7,936	5,290	2,644	0
N_b	0	220	441	661	882
AP	46.0	47.2	**48.3**	47.9	45.1

Interestingly, we verify that heterogeneous combinations can indeed achieve meaningful performance gain. For instance, using a balanced mix of tags and boxes resulted in over 2% improvement compared to using only class tags or only boxes (refer to the bottom row). We believe this gain arises from the complementary information provided by the different types of weak labels, highlighting the potential for further exploration of heterogeneous settings in WSIS.

5.4. Additional Experimental Results

5.4.1 Component Analysis

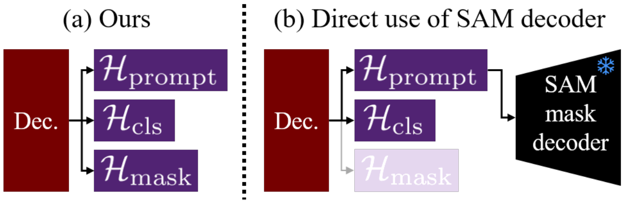
To gain a deeper understanding of the WISH framework, we conducted an ablation study under a homogeneous setting using class tags, as shown in Table 4. First, we analyzed the approach of conditioning segmentation through prompts derived from weak labels and directly predicting these prompts for instance matching. A comparison of Exps A and F demonstrates that the proposed use of $\mathcal{L}_{\text{prompt}}$ yields a significant performance gain. This indicates that the WISH framework effectively learns segmentation through prompt matching. Additionally, we examined the optimal way to utilize the various levels of masks returned by SAM in the mask matching described in Equ. (13). Compared to Exps B-D that uses a specific single level of SAM mask, Exp F shows meaningfully improved performance, by adaptively selecting the mask level that minimizes IoU with the predicted instance mask. Finally, comparing Exps E and F, we observe that $\mathcal{L}_{\text{self}}$ significantly boosts performance. This suggests that localization learned from WISH effectively guides CAMs, enabling the WISH framework to leverage class tags more effectively.

5.4.2 Direct Use of SAM Decoder

In the WISH framework, we define an explicit mask head and use the prompt head solely for matching and loss during training. An interesting alternative is to use the prompt head with the SAM [17] decoder directly for mask prediction [42], as illustrated on the right side of Fig. 3. However, this leads to a significant drop in performance compared to the original (over a 3% decrease in mAP). This was evaluated with a quite lenient criterion, selecting the mask closest to the ground truth among the three masks predicted by the SAM decoder. These results suggest that, rather than simply using SAM as an off-the-shelf module for inference, incorporating SAM's knowledge into the model during training, as done in WISH, provides a more effective approach.

Table 4. Ablation study for WISH framework on PASCAL dataset. We verify the impact of $\mathcal{L}_{\text{prompt}}$, $\mathcal{L}_{\text{mask}}$, and $\mathcal{L}_{\text{self}}$ on performance.

Exp	$\mathcal{L}_{\text{prompt}}$	$\mathcal{L}_{\text{mask}}$	$\mathcal{L}_{\text{self}}$	mAP
A		Adaptive	✓	43.6
B	✓	Lv1	✓	36.2
C	✓	Lv2	✓	40.7
D	✓	Lv3	✓	42.8
E	✓	Adaptive		41.5
F (WISH)	✓	Adaptive	✓	**46.0**


(a) Ours

(b) Direct use of SAM decoder

Dec.
$\mathcal{H}_{\text{prompt}}$
$\mathcal{H}_{\text{cls}}$
$\mathcal{H}_{\text{mask}}$
Dec.
$\mathcal{H}_{\text{prompt}}$
$\mathcal{H}_{\text{cls}}$
$\mathcal{H}_{\text{mask}}$
SAM mask decoder

Figure 3. **(a):** The original WISH framework predicting masks through the mask head, **(b):** An alternative leveraging the prompt head followed by the SAM mask decoder to predict masks.

6. Discussion and Conclusion

In this paper, we introduced WISH, a novel framework for WSIS that leverages a heterogeneous setting, enabling the simultaneous use of multiple types of weak labels. By efficiently integrating diverse weak labels through SAM's prompt latent space and a multi-stage matching strategy, WISH unifies heterogeneous supervision within a single model, outperforming existing homogeneous methods on benchmarks such as Pascal VOC and COCO. Additionally, our proposed protocol allows for exploring optimal annotation strategies, achieving superior performance under heterogeneous settings at equivalent annotation costs, providing valuable insights into practical annotation strategies for WSIS. While the proposed protocol considers annotation costs across different weak label types, it does not account for the costs of image acquisition, nor does it utilize unlabeled images that could serve as additional training data in a semi-supervised manner. Incorporating these factors could enable a more comprehensive evaluation of cost-efficiency and further enhance WISH's applicability to real-world scenarios. Addressing these aspects, especially the inclusion of unlabeled images, represents a promising direction for future work to achieve a fairer comparison and expand the practical utility of WSIS frameworks. In conclusion, we believe that WISH provides a foundational step toward practical and cost-effective WSIS, serving as a strong basis for future research into heterogeneous weak labels and efficient annotation strategies in the broader field.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF2022R1A2B5B03002636) and the MSIT(Ministry of Science and ICT), Korea, under the Graduate School of Metaverse Convergence support program(RS-2024-00418847) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

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, 3, 5, 6

[2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 1, 2

[3] Míriam Bellver Bueno, Amaia Salvador Aguilera, Jordi Torres Viñals, and Xavier Giró Nieto. Budget-aware semi-supervised semantic and instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019*, pages 93–102, 2019. 2

[4] Tianle Chen, Zheda Mai, Ruiwen Li, and Wei-lun Chao. Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. *arXiv preprint arXiv:2305.05803*, 2023. 2

[5] Zhaozheng Chen and Qianru Sun. Weakly-supervised semantic segmentation with image-level labels: from traditional models to foundation models. *arXiv preprint arXiv:2310.13026*, 2023. 2

[6] 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. 1, 3, 4, 5

[7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1, 3, 4, 5, 6, 7

[8] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Qian Zhang, and Wenyu Liu. Boxteacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3145–3154, 2023. 1, 2, 3, 6, 7

[9] Hoonhee Cho, Sung-Hoon Yoon, Hyeokjun Kweon, and Kuk-Jin Yoon. Finding meaning in points: Weakly supervised semantic segmentation for event cameras. In *European Conference on Computer Vision*, pages 266–286. Springer, 2024. 2

[10] 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. 2, 6, 7

[11] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 3, 6, 7

[13] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. *Advances in neural information processing systems*, 28, 2015. 2

[14] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella X Yu. Universal weakly supervised segmentation by pixel-to-segment contrastive learning. *arXiv preprint arXiv:2105.00957*, 2021. 2

[15] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 1, 2, 3

[16] 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, 3, 5, 6, 7

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3, 4, 5, 8

[18] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[19] Hyeokjun Kweon and Kuk-Jin Yoon. Joint learning of 2d-3d weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 35:30499–30511, 2022. 5

[20] Hyeokjun Kweon and Kuk-Jin Yoon. From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19499–19509, 2024. 2, 5

[21] Hyeokjun Kweon, Sung-Hoon Yoon, Hyeonseong Kim, Daehee Park, and Kuk-Jin Yoon. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6994–7003, 2021. 5

[22] Hyeokjun Kweon, Sung-Hoon Yoon, and Kuk-Jin Yoon. Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11329–11339, 2023. 5

[23] Hyeokjun Kweon, Jihun Kim, and Kuk-Jin Yoon. Weakly supervised point cloud semantic segmentation via artificial oracle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3721–3731, 2024. 2

[24] 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, 3, 6, 7

[25] 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*, pages 23745–23755, 2023. 1, 2, 3, 7

[26] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Proposal-based instance segmentation with point supervision. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2126–2130. IEEE, 2020. 1, 2, 3

[27] Hyeonjun Lee, Sehyun Hwang, and Suha Kwak. Extreme point supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17212–17222, 2024. 1, 2, 3, 6

[28] Lei Li, Yongfeng Zhang, and Li Chen. Prompt distillation for efficient llm-based recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1348–1357, 2023. 4

[29] Ruihuang Li, Chenhang He, Yabin Zhang, Shuai Li, Liyi Chen, and Lei Zhang. Sim: Semantic-aware instance mask generation for box-supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7193–7203, 2023. 6, 7

[30] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Xian-Sheng Hua, and Lei Zhang. Box-supervised instance segmentation with level set evolution. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1, 2, 3, 6, 7

[31] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 5

[32] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26617–26626, 2024. 4

[33] Mingxiang Liao, Zonghao Guo, Yuze Wang, Peng Yuan, Bailan Feng, and Fang Wan. Attentionshift: Iteratively estimated part-based attention map for pointly supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19519–19528, 2023. 1, 2, 3, 6, 7

[34] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 6, 7

[35] Bingyuan Liu, Christian Desrosiers, Ismail Ben Ayed, and Jose Dolz. Segmentation with mixed supervision: Confidence maximization helps knowledge distillation. *Medical Image Analysis*, 83:102670, 2023. 2

[36] Kazuya Nishimura and Ryoma Bise. Weakly supervised cell-instance segmentation with two types of weak labels by single instance pasting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3185–3194, 2023. 2

[37] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015. 2

[38] Weixuan Sun, Zheyuan Liu, Yanhao Zhang, Yiran Zhong, and Nick Barnes. An alternative to wsss? an empirical study of the segment anything model (sam) on weakly-supervised semantic segmentation problems. *arXiv preprint arXiv:2305.01586*, 2023. 2

[39] 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 (CVPR)*, pages 1–10, 2021. 1, 2, 3, 6, 7

[40] 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, 3, 7

[41] Zhaoyang Wei, Pengfei Chen, Xuehui Yu, Guorong Li, Jianbin Jiao, and Zhenjun Han. Semantic-aware sam for point-prompted instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2024. 1, 2, 3, 6, 7

[42] Xin-Jian Wu, Ruisong Zhang, Jie Qin, Shijie Ma, and Cheng-Lin Liu. Wps-sam: Towards weakly-supervised part segmentation with foundation models. In *European Conference on Computer Vision*, pages 314–333. Springer, 2025. 2, 4, 8

[43] Sung-Hoon Yoon, Hyeokjun Kweon, Jaeseok Jeong, Hyeonseong Kim, Shinjeong Kim, and Kuk-Jin Yoon. Exploring pixel-level self-supervision for weakly supervised semantic segmentation. *arXiv preprint arXiv:2112.05351*, 2021. 5

[44] Sung-Hoon Yoon, Hyeokjun Kweon, Jegyeong Cho, Shinjeong Kim, and Kuk-Jin Yoon. Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 326–344. Springer Nature Switzerland Cham, 2022. 5

[45] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, Lihe Zhang, Mingyang Qian, and Yizhou Yu. Multi-source weak supervision for saliency detection. In *Proceedings of the IEEE/CVF*

conference on computer vision and pattern recognition, pages 6074–6083, 2019. 2

[46] 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. 2, 5

[47] 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. 1, 2, 3, 5

[48] Lianghui Zhu, Junwei Zhou, Yan Liu, Xin Hao, Wenyu Liu, and Xinggang Wang. Weaksam: Segment anything meets weakly-supervised instance-level recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7947–7956, 2024. 2

[49] 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, 3