WPS-SAM: Towards Weakly-Supervised Part Segmentation with Foundation Models

Xin-Jian Wu[1,2], Ruisong Zhang[1,2], Jie Qin[1,2], Shijie Ma[1,2], Cheng-Lin Liu[1,2⋆]

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
{wuxinjian2020, zhangruisong2019, qinjie2019, mashijie2021}@ia.ac.cn, {liucl}@nlpr.ia.ac.cn

Abstract. Segmenting and recognizing diverse object parts is crucial in computer vision and robotics. Despite significant progress in object segmentation, part-level segmentation remains underexplored due to complex boundaries and scarce annotated data. To address this, we propose a novel **W**eakly-supervised **P**art **S**egmentation (**WPS**) setting and an approach called **WPS-SAM**, built on the large-scale pre-trained vision foundation model, Segment Anything Model (SAM). WPS-SAM is an end-to-end framework designed to extract prompt tokens directly from images and perform pixel-level segmentation of part regions. During its training phase, it only uses weakly supervised labels in the form of bounding boxes or points. Extensive experiments demonstrate that, through exploiting the rich knowledge embedded in pre-trained foundation models, WPS-SAM outperforms other segmentation models trained with pixel-level strong annotations. Specifically, WPS-SAM achieves 68.93% mIOU and 79.53% mACC on the PartImageNet dataset, surpassing state-of-the-art fully supervised methods by approximately 4% in terms of mIOU.

Keywords: Weakly-supervised part segmentation · Foundation models · Part prompts learning

1 Introduction

Recognizing objects and decomposing them into meaningful semantic parts is an inherent capability of human visual perception. This capability is important for human interaction with objects in the physical environment. For instance, when we type on the keyboard, it is necessary to accurately identify the boundaries of keys, as well as understand and discern the functions associated with each key. In perception and reasoning, we frequently deduce the whole from the observation of the parts of an object. This aligns with the conjecture advanced by cognitive psychologists [3, 39], suggesting that hierarchical representations of objects are constructed in a bottom-up manner. Hence, developing a vision system capable

⋆ Corresponding author.

Fig. 1: Illustration of the training data comparison: (a) fully-supervised part segmentation task, (b) proposed WPS task, and (c) WSSS task. Our approach significantly alleviates the burden of data annotation compared to fully-supervised methods, while outperforming WSSS methods in finer-grained tasks.

of part-level segmentation is crucial and promises significant advantages across various applications in computer vision and robotics.

As one of the fundamental tasks in computer vision, object-level semantic segmentation has been extensively studied and has made significant progress [7, 11, 33, 51, 70]. However, part-level segmentation presents additional challenges. Parts often possess intricate structures, characterized by complex boundaries and variations in appearance. Moreover, compared to object segmentation, there is a scarcity of pixel-level annotated data available for training part segmentation models. This scarcity has led to the emergence of various self-supervised and unsupervised part segmentation methods in the early stages [30,46,64]. However, these methods often exhibit lower performance and are only applicable to a limited range of object categories.

To tackle the aforementioned challenges, we propose a novel task called Weakly-supervised Part Segmentation (WPS) in this paper, as illustrated in Figure 1 (b). WPS enables our model to utilize cost-effective annotations, such as points or bounding boxes of parts during the training phase, resulting in pixel-level segmentation of part regions. Compared with existing weakly-supervised semantic segmentation (WSSS) tasks that primarily focus on object-level segmentation with image-level labels [38, 58], WPS specifically addresses the more fine-grained task of part-level semantic segmentation using sparse bounding box-level annotations. Our approach strikes a superior balance between the cost of annotation and the performance of part segmentation.

Empowered by extensive pre-training, vision foundation models like CLIP [55], ALIGN [32], and DINOv2 [52] have exhibited substantial potential in vision understanding, and assisting computer vision tasks, such as enabling cross-modal learning, zero-shot learning, and weakly-supervised learning. Notably, the Segment Anything Model (SAM) [35] has recently emerged, showcasing remarkable segmentation capabilities on unseen objects when dealing with a wide range of images and objects. This provides an opportunity for the part segmentation field.

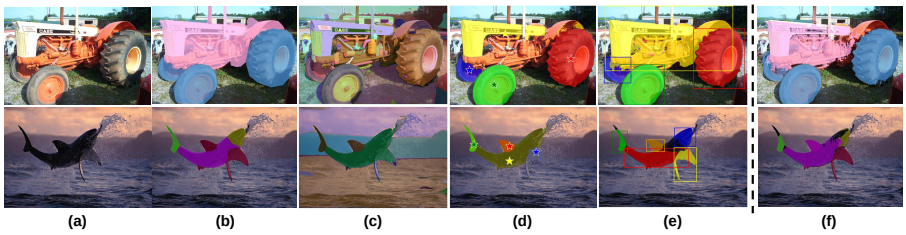


Fig. 2: Visualizations of the segmentation results using pre-trained SAM directly under different modes and employing our method. Each color represents a unique category. (a) Original images. (b) Ground truths of part segmentation. (c) The "everything" mode of SAM without prompts, segments all elements without considering the characteristics of objects and parts. (d) Segmentation results under points-form prompts, which may either miss or over-segment certain parts. (e) Segmentation results with bounding boxes prompts, achieving superior part segmentation results. (f) High-quality segmentation results of the proposed WPS-SAM method without requiring manual provision of prompts.

However, SAM operates as an interactive framework, requiring a high-quality prior prompt (such as a point, box, mask, or text) alongside the input image to generate instance segmentation results. Additionally, it performs category-agnostic segmentation, as shown in Figure 2. These characteristics hinder SAM directly applied to autonomous part segmentation.

To better exploit the enriched knowledge within foundational models and explore their potential ability in the part segmentation task, we propose a novel end-to-end approach, called **WPS-SAM**, which can automate prompts generation to enhance the capabilities of SAM. Unlike most other segmentation methods, the remarkable feature of our approach is that it does not require pixel-level mask annotations. Instead, it solely relies on bounding box or point-level annotations, paired with their respective class labels. The former is used to guide the learning of the prompter, while the latter compensates for the insufficiency of SAM in lacking the category semantics. This helps part segmentation tasks, especially in the situation of limited annotated data. Specifically, we introduce a prompter based on the feature maps extracted by SAM. This prompter autonomously learns the prompt tokens corresponding to different parts in the input image, thereby replacing the previous interactive framework that relied on sequential manual guidance.

In our experiments on different part segmentation datasets, the proposed approach demonstrates superior performance over state-of-the-art methods for object and part segmentation, even though the proposed model relies on only weakly-supervised forms of annotation. For instance, WPS-SAM achieves 68.93% mIOU and 79.53% mACC on the PartImageNet dataset, surpassing state-of-the-art fully supervised methods by 3.69% and 0.69%, respectively. In summary, we make the following contributions in this paper:

- We introduce a novel task: weakly-supervised part segmentation (WPS), aiming to find a balance between the annotation cost and the segmentation performance in the task of part semantic segmentation.
- By exploring the potential of the foundation model, we propose a part segmentation model (WPS-SAM) that can automatically learn part prompts, making the proposed model an end-to-end framework.
- Experimental results show that our WPS-SAM significantly outperforms state-of-the-art part segmentation methods which are trained with fully supervised pixel-level labels.

2 Related Work

2.1 Part Segmentation

Modeling objects in terms of constituent parts has been a persistent challenge in computer vision, with a well-established and extensive research history in this domain. Beginning with the inception of Pictorial Structure [23] introduced in the early 1970s, numerous methods [10,20–22,24,66,74,78] have been introduced to explicitly model the parts and their spatial relationships within the entire object. The deformable part model (DPM) [21] was once considered the most classic work in the field of object detection. These models collectively emphasize that object-part models provide rich representations and enhance the interpretability of prediction.

With the advancement of technology, there's a growing need for a more fine-grained understanding and segmentation of objects at the part level. In the early stages of the deep learning era, the progress of data-driven part segmentation research has been hindered by the absence of extensive datasets containing corresponding part-level mask annotations. Consequently, several self-supervised and unsupervised part segmentation methods [30, 46, 64] emerged. However, these methods exhibit lower performance and are limited to specific classes. Recently, the introduction of datasets incorporating part annotations for common objects, exemplified by PartImageNet [26] and PACO [56], highlights the growing academic attention on this task, prompting further related research [25, 53, 67]. Nonetheless, these methods still rely on pixel-level annotations, incurring high acquisition costs. In this paper, the WPS-SAM we proposed alleviates the model's reliance on strong supervision labels while achieving satisfactory performance.

2.2 Weakly-Supervised Semantic Segmentation

Existing WSSS methods can be roughly grouped into single-stage and multi-stage techniques. The single-stage methods [2, 54, 59, 63, 75] aim to train an end-to-end segmentation models using image-labels. The integration of classification and segmentation in single-stage approaches poses challenges for further optimization of the segmentation model, leading to suboptimal performance in

current methods. On the other hand, the common pipeline of multi-stage methods [1, 37, 38, 58] is to utilize CAMs as initial seed areas to generate pseudo-labels, and then use them to train a segmentation model, which achieves better performance. To improve the quality of pseudo labels initially generated by CAMs, various methods including adversarial erasing [37,38,40,68], saliency guidance [41,73], affinity learning [1,19], contrast learning [17,77] and boundary-aware [43,58] techniques have been proposed.

In recent developments, there a new trend [9,45,62,72] is to take advantage of pre-trained large foundation models [35, 55] to generate high-quality pseudo labels. Although these methods have achieved state-of-the-art performance in the task, they still fall short of the performance exhibited by fully supervised methods. In addition to the conventional setting of WSSS task, several methods [12, 34, 36, 44, 60, 69] introduce the bounding box annotations to generate proper segmentation masks. Compared with these methods, our approach focuses on a more fine-grained task of part-level segmentation using boxes and learns prompts to activate the potential of SAM, eliminating the need for generating pseudo-labels, and achieving better performance than other fully supervised methods.

2.3 Vision Foundation Models

Leveraging extensive pre-training, vision foundational models have attained remarkable success in the field of computer vision. Motivated by the principles of masked language modeling [14, 49] in natural language processing, MAE [27] adopts an asymmetric encoder-decoder structure and employs masked image inpainting to efficiently train scalable vision Transformer models [16]. MAE demonstrates exceptional fine-tuning performance across a range of downstream tasks. CLIP [55] and ALIGN [32] learn image representations from scratch using over a hundred million image-text pairs, demonstrating remarkable zero-shot image classification capabilities.

While most foundation models are designed to extract accessible knowledge from freely available data, the recent SAM method [35] establishes a data engine involving collaborative model development alongside dataset annotation through model-in-the-loop processes. Thanks to pre-training on 1 billion masks and 11 million images, SAM showcases impressive zero-shot, task-agnostic segmentation performance. This has spurred a series of studies [5,6,31,42,47,48,65,79] applying SAM to specific downstream tasks. Motivated by these advancements, our study explores SAM's robust general segmentation capabilities in weakly-supervised part segmentation, aiming to introduce a fresh perspective to readers.

3 Methodology

In this section, we analyze the characteristics of SAM on the part segmentation task. Then we introduce the proposed WPS-SAM for achieving end-to-end part segmentation while relying on weakly-supervised labels only during training. The training and inference techniques will be presented in sequence.

Table 1: The preliminary experimental results of SAM on the PartImageNet *val* set with various types of prompts and backbones, which reflect the performance **upper bound** of our method.

Backbone	Label	mIoU(%)	mACC(%)
ViT-B [15]	point	72.64	90.63
	bbox	91.14	98.19
ViT-L [15]	point	71.14	89.67
	bbox	91.62	98.34
ViT-H [15]	point	71.45	89.95
	bbox	91.80	98.43

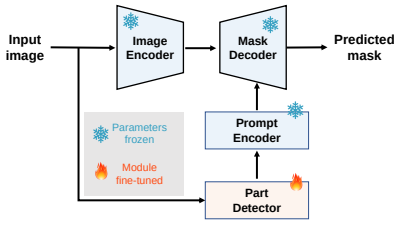


Fig. 3: Schematic diagram of the trivial Det-SAM. We argue that a simple combination of a detector and SAM is not the most optimal solution.

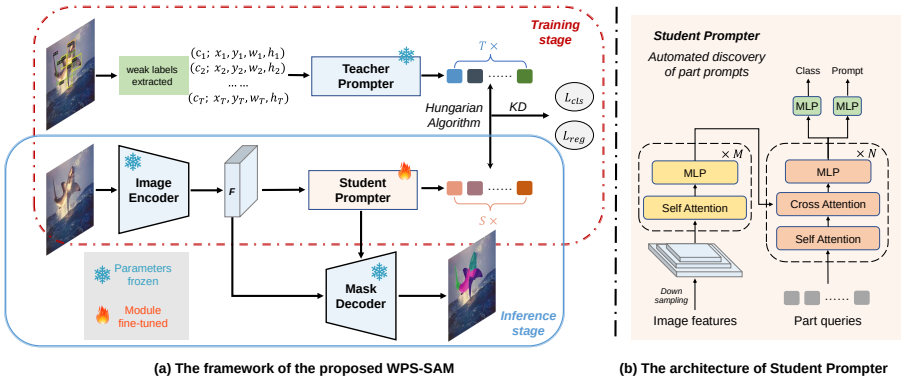
3.1 A Closer Look at SAM

As shown in Figure 2, with high-quality prompts, SAM can yield satisfactory results in part segmentation even without any fine-tuning. To validate this, we conducted preliminary experiments on the PartImageNet dataset with the pre-trained SAM. Specifically, we use annotated bounding boxes and the center points as prompts for SAM input, and the resulting part segmentation outcomes are presented in Table 1. This impressive performance strongly reflects the capabilities of the foundation model in this task. However, manually providing prior prompts is cumbersome in practical applications.

To explore the ability of SAM more efficiently, it is desired to learn or generate prompts automatically. A trivial approach, named Det-SAM in this paper, involves training a dedicated detector initially to detect the bounding boxes of parts. Subsequently, these detected bounding boxes are used as prompts for SAM input, as illustrated in Figure 3. However, we find that this simple strategy does not obtain satisfactory performance because the part detector is imperfect and does not cooperate with the segmentation module very well, more detailed analysis is shown in Section 4.3. To address this issue and further exploit the rich visual information within the foundation model, we propose an end-to-end strategy that directly generates prompts by utilizing the feature maps extracted by the SAM image encoder, which is more efficient and achieves better performance than the above Det-SAM.

3.2 End-to-end Part Segmentation Framework

Overview. As depicted in Figure 4 (a), the overall WPS-SAM architecture is concise and intuitive, consisting of three main modules: the image encoder for feature extraction, the student prompter for prompts generation, and the mask



(a) The framework of the proposed WPS-SAM (b) The architecture of Student Prompter

Fig. 4: An overview of the proposed framework WPS-SAM, accomplishing part segmentation in an end-to-end manner while relying solely on cost-effective weak labels during training. The modules with frozen parameters in the figure come from the pre-trained SAM [35]. Additionally, the utilized student prompts are derived from a lightweight query-based Transformer architecture.

decoder for mask prediction. The entire pipeline is expressed as follows:

$$\begin{aligned} F_{\text{img}} &= \Phi_{\text{i-enc}}(\mathcal{I}) \\ T_{\text{prompt}} &= \Phi_{\text{prompter}}(F_{\text{img}}) \\ \mathcal{M}_{\text{out}} &= \Phi_{\text{m-dec}}(F_{\text{img}}, T_{\text{prompt}}) \end{aligned} \tag{1}$$

where $\mathcal{I} \in \mathbb{R}^{1024\times1024\times3}$ represents the resized original image, $F_{\text{img}} \in \mathbb{R}^{64\times64\times256}$ denotes the embedded image features, and $T_{\text{prompt}} \in \mathbb{R}^{S\times256}$ signifies the prompt tokens encoded by our proposed prompter Φ_{prompter}, $\mathcal{M}_{out} \in \mathbb{R}^{S\times1024\times1024}$ corresponds to the predicted part masks, where S is the number of part queries.

Query-based prompter. As illustrated in Figure 4 (b), the proposed prompter is primarily based on a transformer encoder-decoder structure, incorporating several CNN layers as well as two feed-forward networks.

Considering the large size of the feature map extracted by the image encoder, which imposes a significant computational burden on subsequent modules, we initially designed two layers of 3×3 convolutional layers to implement downsampling and further information fusion.

The encoder plays a crucial role in information fusion and extracting higher-level semantic features from the image features. Subsequently, the decoder is responsible for converting a set of learnable queries into the output embedding by interacting via cross-attention with the semantic features. These output embeddings are then independently predicted into prompt tokens with corresponding categories by the MLPs. These prompt tokens corresponding to the parts in the image are precisely what is needed to accomplish the part segmentation task.

We aim to augment the capabilities of the SAM prompter, enabling it to autonomously derive semantic prompts based on the current input. To preserve the knowledge already acquired by SAM and minimize computational costs, We freeze the parameters in image encoder $\Phi_{\text{i-enc}}$ and mask decoder $\Phi_{\text{m-dec}}$ of SAM. Then we learn the proposed student prompter based on the knowledge distillation technology [29]. Specifically, we employ the original SAM's prompt encoder as the teacher network, utilizing it to receive weakly supervised data annotations, *e.g.*, points or bounding boxes. Subsequently, we use its outputs to supervise the training of the student prompter Φ_{prompter}, which takes image features embedded by $\Phi_{\text{i-enc}}$ as input.

Finally, the trained Φ_{prompter} can replace the original prompter, enabling the end-to-end part segmentation task.

3.3 Training and Inference

Training. The query-based prompter generates a fixed-size set of S predictions, whereas the actual number of parts in the current image T is variable. One of the primary challenges in training is to evaluate the predicted parts (category, prompt embedding) of the pseudo-labels generated by the teacher prompter. Our loss function establishes an optimal bipartite matching between predictions and pseudo-labels and subsequently optimizes parts-specific losses, which simultaneously considers category prediction and the regression task for teacher prompt embeddings.

Let y represent the supervised label set of parts, and $\hat{y} = \{\hat{y}_i\}_{i=1}^{S}$ denote the set of S predictions. Assuming that S is greater than the number of parts in the image, we treat y as a set of size S padded with $\varnothing$ (no parts).

To establish a bipartite matching between these two sets, we seek a permutation of S elements $\sigma \in \sigma_S$ with the minimum cost:

$$\hat{\sigma} = \arg\min_{\sigma \in \sigma_S} \sum_{i}^{S} C_m(y_i, \hat{y}_{\sigma(i)}), \tag{2}$$

where $C_m(y_i, \hat{y}_{\sigma(i)})$ represents the pairwise matching cost between the ground truth y_i and a prediction with index $\sigma(i)$. The optimal assignment is efficiently computed using the Hungarian algorithm, as outlined in prior works [4, 61].

The matching cost considers both the class prediction and the similarity between predicted prompts and pseudo-labels generated by the teacher prompter. Each element i in the supervised label set can be represented as $y_i = (c_i, p_i)$, where c_i is the target class label (which may be $\varnothing$), and p_i is the prompt embedding generated by the teacher prompter based on weak labels. For the prediction with index $\sigma(i)$, we define the probability of class c_i as $\hat{p}_{\sigma(i)}(c_i)$ and the predicted prompts as $\hat{p}_{\sigma(i)}$. With these notations, we define the matching cost as follows:

$$C_m(y_i, \hat{y}_{\sigma(i)}) = \mathbb{1}_{\{c_i \neq \varnothing\}}(-\alpha \hat{p}_{\sigma(i)}(c_i) + \beta L_2(p_i, \hat{p}_{\sigma(i)})), \tag{3}$$

where α and β are weight coefficients used to balance the two types of costs.

Once each predicted instance is paired with its corresponding ground truth under the optimal assignment $\hat{\sigma}$ computed in step 2, we can compute the loss function to optimize the parameters as described below:

$$\mathcal{L}(y, \hat{y}) = \frac{1}{S} \sum_{i}^{S} (\lambda_{cls} \mathcal{L}_{\text{cls}}^{i} + \mathbb{1}_{\{c_i \neq \varnothing\}} \lambda_{reg} \mathcal{L}_{\text{reg}}^{i}), \tag{4}$$

where $\mathcal{L}_{\text{cls}}$ denotes the cross-entropy loss computed between the predicted category and the target (contain the category of $\varnothing$), while $\mathcal{L}_{reg}$ signifies the smooth L_1 loss between the predicted prompt embeddings and the matched teacher prompt embeddings.

Inference. In the inference phase, we no longer rely on the Hungarian matching process, because of the lack of supervised teacher labels. Instead, we directly retain the prompt tokens corresponding to the foreground (specific part categories) and discard those related to the background (empty category $\varnothing$) based on the predictions from the classification head in the student prompter.

4 Experiments

4.1 Setup

This paper aims to achieve weakly-supervised part semantic segmentation. To validate the concept of our problem and the effectiveness of our approaches, we use the following datasets. During the model training, we exclusively utilize part annotations in the form of points or bounding boxes, reserving pixel-level masks for performance evaluation during inference. We adopt standard evaluation metrics for semantic segmentation, *i.e.*, mean Intersection over Union (mIoU), and mean Pixel Accuracy (mACC).

PartImageNet [26]. This dataset is a large, high-quality dataset with part segmentation annotations following the COCO style. It comprises 158 classes from ImageNet [13], totaling 24,080 images. The classes are organized into 11 super-categories, and the part splits are designed based on these super-categories, totaling 40 part categories.

PASCAL-Part [10]. The original Pascal Part dataset offers part annotations for 20 classes from Pascal VOC [18], encompassing a total of 193 part categories. The training and validation sets comprise 10,103 images, while the testing set contains 9,637 images. Following the setting of [25], we only consider the 16 classes that have part-level annotations and ignore the rest. We manually merge the provided labels to a higher-level definition of parts (e.g. "eyes", "ears", "nose", etc. can be merged into a single "head" part) since the original parts are too fine-grained.

Implementation details. During training, we maintain the image size at 1024 $\times$ 1024, consistent with SAM, and refrain from applying additional augmentations. We only train the parameters of the introduced prompter while freezing the parameters of other parts of the network. The student prompter includes a downsampling layer with two layers of 3×3 convolutional kernels with a stride of

Table 2: Comparisons between WPS-SAM and classical fully-supervised segmentation models, as well as state-of-the-art weakly supervised semantic segmentation (WSSS) methods on the **PartImageNet** *val* set.

	Method	Venue	Backbone	Annotation	mIoU(%)	mACC(%)
Fully-Supervised	Deeplab v3+ [8]	ECCV'18	ResNet-50 [28]	mask	60.57	71.07
	MaskFormer [11]	NeurIPS'21			60.34	72.75
	MaskFormer-Dual				58.02	70.42
	Compositor [25]	CVPR'23			61.44	73.41
	SegFormer [70]	NeurIPS'21	MiT-B2 [70]	mask	61.97	73.77
	MaskFormer [11]	NeurIPS'21	Swin-T [50]		63.96	77.37
	MaskFormer-Dual				61.69	75.64
	Compositor [25]	CVPR'23			64.64	78.31
	MaskFormer [11]	NeurIPS'21	Swin-B [50]		65.24	78.84
WSSS	**WPS-SAM**	This Work	ViT-B [15]	bbox	**68.93**	**79.53**
	SIM [44]	CVPR'23	ResNet-101 [28]		49.51	63.27
	BECO [58]	CVPR'23	ResNet-101 [28]	category	42.37	53.07
	FMA-WSSS [72]	WACV'24			56.74	68.07

2. Both the encoder and decoder layers in the student prompter are set to 6 layers. And we set $\alpha = 10.0,\ \beta = 1.0,\ \lambda_{\text{cls}} = 5.0,\ \lambda_{\text{reg}} = 20.0$ respectively. We utilize the Adam optimizer with a learning rate of $1e-4$ for prompter training and set the batch size to 8 on each GPU. The total training epochs amount to 150. Our source code is publicly available at `https://github.com/xjwu1024/WPS-SAM`.

4.2 Main Results

Comparisons with other methods. We conducted comparisons between our proposed WPS-SAM and classical fully-supervised segmentation models, as well as state-of-the-art weakly supervised semantic segmentation (WSSS) methods. The fully-supervised segmentation models contain CNN-based semantic segmentation models [8, 11], as well as more advanced Transformer-based architectures [11, 70]. Additionally, we compared our approach with the latest state-of-the-art research [25] about part segmentation. Even though only sparse annotations at the bounding box level are introduced in our method, the experimental results in Table 2 and Table 3 demonstrate that our method outperforms other methods that use stronger pixel-level dense annotations. On another hand, our approach achieves significantly better performance than existing WSSS methods without the need for cumbersome steps such as generating pseudo-labels. Moreover, leveraging a pre-trained visual foundation model in our approach results in a significant portion of parameters being frozen. This significantly reduces the number of trainable parameters in comparison to alternative methods which leads to a decrease in computational costs during the training process.

Table 2 presents a summary of our experimental results on PartImageNet. Utilizing ViT-B as the backbone, our proposed method, WPS-SAM, achieves 68.93% mIoU and 79.53% mACC. These results surpass those of fully-supervised methods of comparable scale by approximately 4% and state-of-the-art WSSS method by around 12% in terms of mIOU.

Table 3: Comparisons between WPS-SAM and classical fully-supervised segmentation models, as well as state-of-the-art weakly supervised semantic segmentation (WSSS) methods on the **PASCAL-Part** *val* set.

	Method	Venue	Backbone	Annotation	mIoU(%)	mACC(%)
Fully-Supervised	MaskFormer [11]	NeurIPS'21	ResNet-50 [28]	mask	47.61	58.59
	MaskFormer-Dual				46.60	57.96
	Compositor [25]	CVPR'23			48.01	58.83
	MaskFormer [11]	NeurIPS'21	Swin-T [50]	mask	55.42	67.21
	MaskFormer-Dual				54.21	66.42
	Compositor [25]	CVPR'23			55.92	67.63
	MaskFormer [11]	NeurIPS'21	Swin-B [50]		56.83	68.46
WSSS	**WPS-SAM**	This Work	ViT-B [15]	bbox	**60.49**	**71.25**
	SIM [44]	CVPR'23	ResNet-101 [28]		37.82	49.56
	BECO [58]	CVPR'23	ResNet-101 [28]	category	34.53	46.66
	FMA-WSSS [72]	WACV'24			42.21	54.13

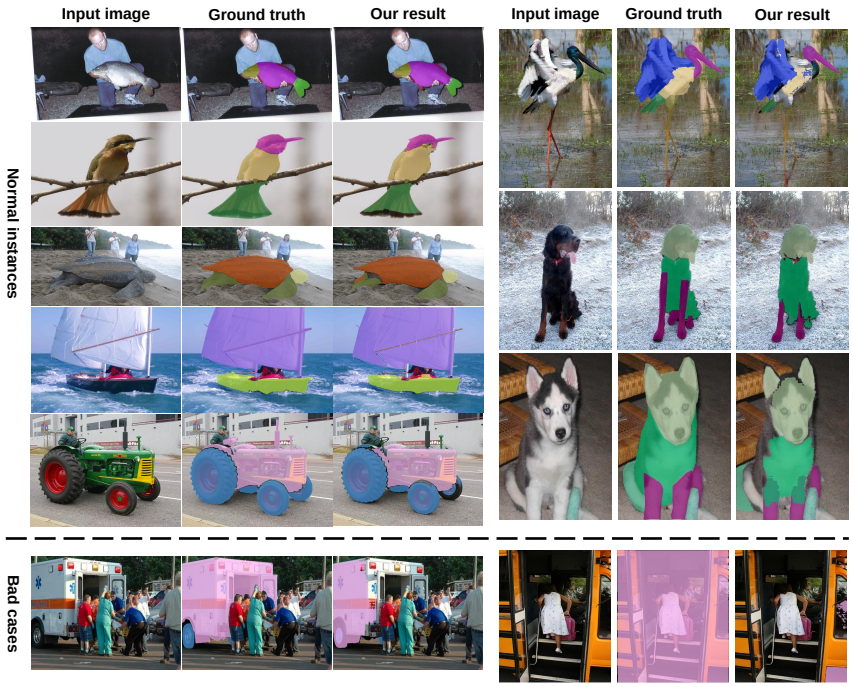


Fig. 5: Qualitative results for part segmentation on PartImageNet generated by WPS-SAM with ViT-B backbone. The masks with different colors correspond to different part categories. These high-quality part predictions demonstrate the feasibility of part segmentation in diverse real-world scenarios and underscore the effectiveness of our proposed approach. We present segmentation results for the majority of typical examples (above the dashed line) and showcase a few bad cases (below the dashed line) where parts are occluded or lack clear semantic information.

Table 4: Performance comparison of different frameworks, where ***Backbone*** refers to the part detector's underlying architecture.

Framework	Detector	Backbone	mAP(%)	mIoU(%)	mACC(%)
Det-SAM	Faster-RCNN [57]	ResNet-50	35.0	63.7	74.8
	DETR [4]		34.2	61.4	70.2
	Faster-RCNN [57]	SAM-Encoder	24.2	54.6	65.2
	DETR [4]		26.3	56.3	68.6
WPS-SAM	-	-	-	**68.9**	**79.5**

Table 3 presents our experimental results on Pascal-Part. Notably, in contrast to the images in PartImageNet, which typically contain a single object, Pascal-Part scenes are more complex, involving multiple objects. This complexity results in a performance decline for various methods, however, our approach maintains a significant advantage. With ViT-B as the backbone, WPS-SAM achieves 60.49% mIoU and 71.25% mACC. In short, we demonstrate that our approach attains stronger performance by leveraging the rich knowledge within SAM, even with weaker annotations.

Visualizations of part segmentation results. In order to offer a more intuitive understanding of the proposed WPS-SAM, we conducted the following visualization analysis, as illustrated in Figure 5. Despite being trained exclusively on sparse, box-level part annotations, our model demonstrates a remarkable capability to generate high-quality part segmentation results that exhibit close alignment with the ground truths. This noteworthy accomplishment underscores the robustness and efficacy of WPS-SAM in capturing intricate part details based on minimal annotation information. Moreover, it is crucial to acknowledge the presence of challenging cases in specific scenarios, such as situations where the target is obscured or semantic information is unclear. This also provides a direction for us to further refine our approach.

4.3 Ablation Study

In this section, we conduct a series of experiments on PartImageNet to investigate the importance of each component and parameter setting in our proposed method. Unless otherwise specified, the ViT-B is used as the default backbone of the image encoder.

Det-SAM *V.S.* WPS-SAM. We evaluate the effectiveness of our framework by considering Det-SAM, which involves training a dedicated part detector independently and connecting it to the prompter of SAM, as discussed in Section 3.1. We utilize two classic detectors, Faster R-CNN [57] and DETR [4], as part detectors and compare their performance with our approach. The results are presented in Table 4. Det-SAM achieves performance comparable to existing methods but significantly lower than our proposed framework, further emphasizing the effectiveness of our approach. We attribute the superior performance of our method to two main factors. Firstly, using a pre-trained image feature extractor, which

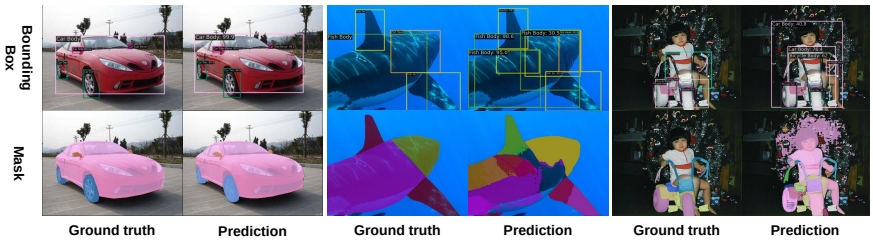


Fig. 6: The visualization results of Det-SAM, with Faster-RCNN as a detector, which includes both box-level and mask-level information, clearly demonstrate the significant impact of part detection on the performance of part segmentation.

Table 5: Performance comparison of MaskFormer with ground truth mask, pseudo-labels, and WPS-SAM on **PartImageNet** *val* set.

Method	Backbone	Supervision	mIoU(%)	mACC(%)
MaskFormer [11]	Swin-B	gt-mask	65.24	78.84
		bbox → pseudo-mask	56.83	69.57
WPS-SAM	ViT-B	bbox	**68.93**	**79.53**

has been trained on large-scale datasets, enables our framework to leverage richer visual information. Secondly, we employ high-dimensional prompt embeddings as teacher labels for training the student prompter, which encapsulates more informative cues than the more traditional bounding box representations. Furthermore, we conducted experiments to evaluate the performance of Det-SAM by replacing the backbone of the part detectors with the pre-trained SAM-encoder. The results showed a decrease in performance, which could be attributed to compatibility issues between the frameworks and suboptimal adjustments of training details. Figure 6 illustrates the qualitative segmentation results obtained using Det-SAM. It is evident that the accuracy of the part detector heavily influences the segmentation performance.

Compared with the pseudo-label paradigm. Based on the preliminary results presented in Table 1, we observed that SAM demonstrates the capability to generate high-quality segmentation masks. These masks can be used as pseudo-labels to train existing segmentation models, following a two-phase approach commonly employed in WSSS methods. Furthermore, we report the performance of MaskFormer trained using these pseudo-labels in Table 5. It is evident that the performance of MaskFormer trained with SAM-generated pseudo-labels is competitive with ground truth masks supervised. However, it still falls slightly behind our WPS-SAM, highlighting the advantages of SAM in the segmentation task compared to other models.

Boxes or Centers? We also explore another widely used form of weak labeling, specifically *center point* annotation, to assess the performance of WPS-SAM. Table 6 demonstrates that under the point annotation form, the performance of WPS-SAM is significantly inferior to that of the bounding box annotation. We attribute this disparity to the fact that while point annotation can locate parts, it fails to capture the shape of the parts. In contrast, bounding boxes can to some extent reflect this information, resulting in stronger performance, as illustrated in Figure 2.

Table 6: Performance comparison on points and boxes supervised.

supervision	mIoU(%)	mACC(%)
point	52.28	66.85
bbox	**68.93**	**79.53**

Number of queries. Considering that the number of parts in a single image is usually not too large (generally not exceeding 20), so we set the number of queries in our proposed prompter to 25. Table 7 presents the performance of the model under different query quantities. When increasing the parameter count to 50/75/100, we noticed a slight decrease in performance. This decrease may be attributed to the additional convergence burden placed on the model due to the larger parameter count.

Table 7: Ablation study on the number of queries.

#queries	mIoU(%)	mACC(%)
25	**68.93**	**79.53**
50	64.74	75.07
75	62.93	72.40
100	62.71	71.80

Due to space limitations, more discussions can be seen in the appendix.

5 Conclusion

In this paper, we introduce WPS-SAM, a novel weakly-supervised prompt learning method for part segmentation. WPS-SAM strikes a balance between annotation cost and segmentation performance by automatically learning part prompts, without manual guidance. Leveraging pre-trained foundation models, WPS-SAM outperforms other segmentation methods relying on strong-supervised annotations, achieving remarkable results with a 68.93% mIOU and 79.53% mACC on the PartImageNet dataset. We conduct a comprehensive analysis to highlight the advantages of our approach over Det-SAM and the pseudo-label paradigm. Additionally, we delve into an exploration of the performance of our method under different weak supervision (box and point), along with an investigation into the impact of various hyperparameters on our model. Despite the promising results, a limitation of our proposed method is its reliance on SAM, which is computationally heavy and hinders its applicability in resource-constrained environments. We suggest incorporating lightweight variants of SAM. such as EfficientSAM [71] and FastSAM [76], to mitigate this limitation.

Overall, our work offers a valuable contribution to the development of part segmentation, showcasing the effectiveness of incorporating foundation models. We hope that our method inspires new research and leads to further improvements in weakly-supervised part segmentation.

Acknowledgements

This work has been supported by the National Natural Science Foundation of China (NSFC) grant U20A20223.

References

1. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4981–4990 (2018)
2. Araslanov, N., Roth, S.: Single-stage semantic segmentation from image labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4253–4262 (2020)
3. Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychological Review **94**(2), 115 (1987)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
5. Chen, J., Yang, Z., Zhang, L.: Semantic segment anything (2023)
6. Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., Shi, Z.: Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. arXiv preprint arXiv:2306.16269 (2023)
7. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
8. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
9. Chen, T., Mai, Z., Li, R., Chao, W.l.: Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. arXiv preprint arXiv:2305.05803 (2023)
10. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.: Detect what you can: Detecting and representing objects using holistic models and body parts. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1971–1978 (2014)
11. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems **34**, 17864–17875 (2021)
12. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1635–1643 (2015)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)

16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
17. Du, Y., Fu, Z., Liu, Q., Wang, Y.: Weakly supervised semantic segmentation by pixel-to-prototype contrast. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4320–4329 (2022)
18. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**, 303–338 (2010)
19. Fan, J., Zhang, Z., Tan, T., Song, C., Xiao, J.: Cian: Cross-image affinity net for weakly supervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10762–10769 (2020)
20. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE transactions on pattern analysis and machine intelligence **28**(4), 594–611 (2006)
21. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence **32**(9), 1627–1645 (2009)
22. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. International journal of computer vision **61**, 55–79 (2005)
23. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Transactions on computers **100**(1), 67–92 (1973)
24. Girshick, R., Felzenszwalb, P., McAllester, D.: Object detection with grammar models. Advances in neural information processing systems **24** (2011)
25. He, J., Chen, J., Lin, M.X., Yu, Q., Yuille, A.L.: Compositor: Bottom-up clustering and compositing for robust part and object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11259–11268 (2023)
26. He, J., Yang, S., Yang, S., Kortylewski, A., Yuan, X., Chen, J.N., Liu, S., Yang, C., Yu, Q., Yuille, A.: Partimagenet: A large, high-quality dataset of parts. In: European Conference on Computer Vision. pp. 128–145. Springer (2022)
27. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
29. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
30. Hung, W.C., Jampani, V., Liu, S., Molchanov, P., Yang, M.H., Kautz, J.: Scops: Self-supervised co-part segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 869–878 (2019)
31. Ji, W., Li, J., Bi, Q., Liu, T., Li, W., Cheng, L.: Segment anything is not always perfect: An investigation of sam on different real-world applications. Machine Intelligence Research **21**, 1–14 (2024)
32. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021)

33. Jiang, R., Zhu, R., Su, H., Li, Y., Xie, Y., Zou, W.: Deep learning-based moving object segmentation: Recent progress and research prospects. Machine Intelligence Research **20**, 335–369 (2023)
34. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 876–885 (2017)
35. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
36. Kulharia, V., Chandra, S., Agrawal, A., Torr, P., Tyagi, A.: Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In: European Conference on Computer Vision. pp. 290–308. Springer (2020)
37. Kweon, H., Yoon, S.H., Kim, H., Park, D., Yoon, K.J.: 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. pp. 6994–7003 (2021)
38. Kweon, H., Yoon, S.H., Yoon, K.J.: Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11329–11339 (2023)
39. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015)
40. Lee, J., Kim, E., Yoon, S.: Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4071–4080 (2021)
41. Lee, S., Lee, M., Lee, J., Shim, H.: Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5495–5505 (2021)
42. Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Yang, J., Li, C., Zhang, L., Gao, J.: Semantic-sam: Segment and recognize anything at any granularity. arXiv preprint arXiv:2307.04767 (2023)
43. Li, J., Fan, J., Zhang, Z.: Towards noiseless object contours for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16856–16865 (2022)
44. Li, R., He, C., Zhang, Y., Li, S., Chen, L., Zhang, L.: Sim: Semantic-aware instance mask generation for box-supervised instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7193–7203 (2023)
45. Lin, Y., Chen, M., Wang, W., Wu, B., Li, K., Lin, B., Liu, H., He, X.: Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15305–15314 (2023)
46. Liu, S., Zhang, L., Yang, X., Su, H., Zhu, J.: Unsupervised part segmentation through disentangling appearance and shape. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8355–8364 (2021)
47. Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., Shen, C.: Matcher: Segment anything with one shot using all-purpose feature matching. arXiv preprint arXiv:2305.13310 (2023)

48. Liu, Y., Zhang, J., She, Z., Kheradmand, A., Armand, M.: Samm (segment any medical model): A 3d slicer integration to sam. arXiv preprint arXiv:2304.05622 (2023)
49. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
50. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
51. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)
52. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
53. Pan, T.Y., Liu, Q., Chao, W.L., Price, B.: Towards open-world segmentation of parts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15392–15401 (2023)
54. Pathak, D., Krahenbuhl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1796–1804 (2015)
55. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
56. Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., Wang, R., Marquez, A., Kovvuri, R., Kadian, A., et al.: Paco: Parts and attributes of common objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7141–7151 (2023)
57. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28 (2015)
58. Rong, S., Tu, B., Wang, Z., Li, J.: Boundary-enhanced co-training for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19574–19584 (2023)
59. Ru, L., Zhan, Y., Yu, B., Du, B.: Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16846–16855 (2022)
60. Song, C., Huang, Y., Ouyang, W., Wang, L.: Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3136–3145 (2019)
61. Stewart, R., Andriluka, M., Ng, A.Y.: End-to-end people detection in crowded scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2325–2333 (2016)
62. Sun, W., Liu, Z., Zhang, Y., Zhong, Y., Barnes, N.: 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)

63. Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C., Boykov, Y.: On regularized losses for weakly-supervised cnn segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 507–522 (2018)
64. Thewlis, J., Bilen, H., Vedaldi, A.: Unsupervised learning of object landmarks by factorized spatial embeddings. In: Proceedings of the IEEE international conference on computer vision. pp. 5916–5925 (2017)
65. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Segmenting everything in context. arXiv preprint arXiv:2304.03284 (2023)
66. Weber, M., Welling, M., Perona, P.: Unsupervised learning of models for recognition. In: Computer Vision-ECCV 2000: 6th European Conference on Computer Vision Dublin, Ireland, June 26–July 1, 2000 Proceedings, Part I 6. pp. 18–32. Springer (2000)
67. Wei, M., Yue, X., Zhang, W., Kong, S., Liu, X., Pang, J.: Ov-parts: Towards open-vocabulary part segmentation. arXiv preprint arXiv:2310.05107 (2023)
68. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1568–1576 (2017)
69. Xie, C., Ren, D., Wang, L., Zuo, W.: Learning class-agnostic pseudo mask generation for box-supervised semantic segmentation. arXiv preprint arXiv:2103.05463 (2021)
70. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
71. Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., et al.: Efficientsam: Leveraged masked image pretraining for efficient segment anything. arXiv preprint arXiv:2312.00863 (2023)
72. Yang, X., Gong, X.: Foundation model assisted weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 523–532 (2024)
73. Yao, Q., Gong, X.: Saliency guided self-attention network for weakly and semi-supervised semantic segmentation. IEEE Access **8**, 14413–14423 (2020)
74. Yuille, A.L., Hallinan, P.W., Cohen, D.S.: Feature extraction from faces using deformable templates. International journal of computer vision **8**, 99–111 (1992)
75. Zhang, B., Xiao, J., Wei, Y., Sun, M., Huang, K.: Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12765–12772 (2020)
76. Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., Wang, J.: Fast segment anything. arXiv preprint arXiv:2306.12156 (2023)
77. Zhou, T., Zhang, M., Zhao, F., Li, J.: Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4299–4309 (2022)
78. Zhu, S.C., Mumford, D., et al.: A stochastic grammar of images. Foundations and Trends® in Computer Graphics and Vision **2**(4), 259–362 (2007)
79. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. arXiv preprint arXiv:2304.06718 (2023)