Point2Mask: Point-supervised Panoptic Segmentation via Optimal Transport

Wentong Li[1], Yuqian Yuan[1], Song Wang[1],
Jianke Zhu[1]*, Jianshu Li[2], Jian Liu[2], Lei Zhang[3]
[1]Zhejiang University [2]Ant Group [3]The HongKong Polytechnical University

Figure 1: **Examples of pixel-wise mask predictions generated by Point2Mask** on COCO with ResNet-101. Only a *single point* annotation per target is used as supervision during training to obtain these results.

Abstract

Weakly-supervised image segmentation has recently attracted increasing research attentions, aiming to avoid the expensive pixel-wise labeling. In this paper, we present an effective method, namely Point2Mask, to achieve high-quality panoptic prediction using only a single random point annotation per target for training. Specifically, we formulate the panoptic pseudo-mask generation as an Optimal Transport (OT) problem, where each ground-truth (gt) point label and pixel sample are defined as the label supplier and consumer, respectively. The transportation cost is calculated by the introduced task-oriented maps, which focus on the category-wise and instance-wise differences among the various thing and stuff targets. Furthermore, a centroid-based scheme is proposed to set the accurate unit number for each gt point supplier. Hence, the pseudo-mask generation is converted into finding the optimal transport plan at a globally minimal transportation cost, which can be solved via the Sinkhorn-Knopp Iteration. Experimental results on Pascal VOC and COCO demonstrate the promising performance of our proposed Point2Mask approach to point-supervised panoptic segmentation. Source code is available at: https://github.com/LiWentomng/Point2Mask.

1. Introduction

Panoptic segmentation aims to obtain the pixel-wise labels of instance things and semantic stuff in the whole image, which plays an important role in applications such as autonomous driving, image editing and robotic manipulation. Although having achieved promising performance, most of the existing panoptic segmentation approaches [29, 10, 51, 8, 20, 47] are trained in a fully supervised manner, which heavily depend on the pixel-wise mask annotations, incurring expensive labeling costs.

To deal with this problem, weakly-supervised methods have recently attracted research attentions to obtain high-quality pixel-wise masks with label-efficient sparse annotations, such as bounding box [43, 26, 22, 27], multiple points [28], or the combination of them [9, 41]. Such methods make image segmentation more accessible with lower annotation efforts for new categories or scene types. In this paper, we explore a simpler yet more efficient annotation form, *i.e.*, *a single random point* for each thing and stuff target, to achieve high-quality panoptic segmentation. As discussed in [2], the cost of point-level labels is only marginally above image-level ones [1]. Such a setting has

[1]On Pascal VOC [14], image labels cost around 20 sec./img, single point labels cost 22.1 sec./img, while full mask labels cost 239.7 sec./img.

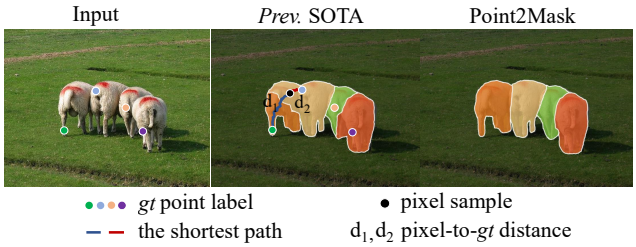

Figure 2: By taking an image with a single random gt point label per target as the input, the method in [15] adopts the minimum distance for each pixel-gt pair to determine the pseudo label, which cannot handle the ambiguous locations and heavily relies on the defined distance. For example, d_2 is shorter than d_1 for the current pixel in black color, which results in wrong assignment. Our Point2Mask formulates this task as a global Optimal Transport problem, and obtains accurate pseudo-mask labels.

been rarely studied due to the little available supervision information from a single point for pixel-wise mask prediction. Only one recent study [15] has attempted to build the minimum traversing distance between each pair of pixel sample and ground-truth (denoted as gt) point label to determine the accurate pseudo mask label.

Unfortunately, it is sub-optimal to assign the pixel samples independently for each random gt point label according to the defined minimum distance. As shown in Fig. 2, the previous method [15] heavily relies on the defined distance and lacks the global context in dealing with the ambiguous locations (*i.e.*, the border pixels among different thing-based targets with the same category). The pixel-to-gt assignment for ambiguous samples is non-trivial, which requires further information beyond the local view. To this end, we model this task from a global optimization perspective to determine the high-quality pixel sample partition for all gt point labels within an image.

In this paper, we propose a novel single point-supervised panoptic segmentation method, dubbed as *Point2Mask*, which formulates the pseudo-mask generation as an Optimal Transport (OT) problem. Specifically, we firstly define each gt point label as a supplier who provides a certain number of labels, and regard each pixel sample as a consumer who needs one unit gt label. To accurately define the transportation cost between each pixel-gt pair, we introduce two types of task-oriented maps, including category-wise semantic map and instance-wise boundary map. The former focuses on the semantic differences among the categories, while the later aims to discriminate the thing-based objects with accurate boundary. Furthermore, we propose an effective centroid-based scheme to set the accurate unit number for each gt point supplier in the OT problem.

Under our proposed framework, the pseudo-mask generation is converted into finding the optimal transport plan at a globally minimal transportation cost, which can be efficiently solved via the Sinkhorn-Knopp Iteration [12]. By making use of the pseudo-mask labels, the panoptic segmentation sub-network is optimized in a fully-supervised manner. The proposed Point2Mask method is an end-to-end training framework, where only the fully-supervised sub-network is retained for inference. Extensive experiments are conducted on Pascal VOC [14] and COCO [31] benchmarks, and the promising qualitative and quantitative results demonstrate the effectiveness of our proposed approach. Notably, Point2Mask surpasses the state-of-the-art method [15] by 4.0% PQ on Pascal VOC and 3.1% PQ on COCO with the same ResNet-50 backbone [18], and achieves comparable performance with the fully-supervised methods using the Swin-L backbone [32]. Some qualitative results are shown in Fig. 1.

2. Related Work

Fully-supervised Panoptic Segmentation. Image segmentation tackles the problem of grouping pixels. As the unified image segmentation task, panoptic segmentation [21] simultaneously incorporates semantic and instance segmentation, where each pixel is uniquely assigned with one of the stuff classes or one of the thing instances.

To this end, some methods [21, 45, 7] have been proposed by dealing with things and stuff using separate network branches within one model. Recently, some works [29, 10, 44, 51, 8, 23] aim to unify the model for this task. DETR [3] predicts the boxes for things and stuff categories with Transformer to perform panoptic segmentation. Mask2Former [8] further employs an additional pixel decoder to take into account of the high-resolution features and generates the mask predictions by the Transformer decoder with the masked-attention. Despite being able to segment objects with accurate boundaries, these methods rely on the expensive and laborious pixel-wise mask annotations, which hinders them from dealing with new categories or scene types in real-world applications [37, 46, 49, 48, 5].

Weakly-supervised Panoptic Segmentation. Weakly supervised segmentation intends to alleviate the annotation burden in segmentation tasks by label-efficient sparse labels for training. According to different kinds of tasks, it ranges from semantic segmentation [50, 30, 19, 42] to instance segmentation [9, 43, 22, 26, 27, 1] and to panoptic segmentation [15, 38, 28] tasks. As for panoptic segmentation, Li *et al.* [28] employed coarse polygons with multiple point annotations for each target to supervise the panoptic segmentation model. Recently, Fan *et al.* [15] adopted a simpler labeling form, *i.e.*, a single point annotation, for each target in an image, and introduced the minimum traversing distance between each pixel sample and the target point label. In spite of its promising performance, it heavily relies on the defined distance, which cannot handle the ambiguous

border locations with a local view. Thus, it is still challenging to obtain the accurate mask predictions for single point-supervised panoptic segmentation.

Optimal Transport in Computer Vision. The Optimal Transport (OT) is a classical optimization problem with a wide range of computer vision applications. In the early years, the Wasserstein distance (WD), also known as the Earth Mover's distance, was adopted to capture the structure of color distribution and texture spaces for image retrieval [35]. Recently, Chen *et al.* [6] employed OT to explicitly encourage the fine-grained alignment between words and image regions for vision-and-language pre-training. Li *et al.* [24] built an attention-aware transport distance in OT to measure the discriminant information from domain knowledge for unsupervised domain adaptation. To achieve high-quality label assignment, Ge *et al.* [16] formulated the label assignment in object detection as the problem of solving an OT plan. In this work, we explore OT for point-supervised panoptic segmentation.

3. Method

3.1. Overview of Point2Mask

As illustrated in Fig. 3, we leverage a unified framework, namely Point2Mask, for single point-supervised panoptic segmentation. It consists of two network branches. One branch generates the mask pseudo-labels, and the other focuses on the fully supervised learning using Panoptic SegFormer model [29] based on the generated pseudo-labels. The two branches share the basic backbone and neck network, which are trained in an end-to-end fashion. The key of our proposed approach is how to model the process of mask pseudo-label generation as the global Optimal Transport (OT) problem, which aims to obtain the accurate pixel-wise pseudo-masks with only a single point label per target.

3.2. Optimal Transport

We first give a brief review of OT [34], which aims to find a transportation plan Γ minimizing the total cost of moving goods from one location to another. It is subject to certain constraints on the amount of goods to be transported and the cost of transportation.

Given a set of m suppliers, another set of n consumers, and a cost function c_{ij} that specifies the cost of transporting one unit of goods from the i-th supplier to the j-th consumer. The goal of OT is to find a transportation plan $\Gamma = \{\Gamma_{i,j} \,|i = 1, 2, \cdots, m, j = 1, 2, \cdots, n\}$ that minimizes the total cost of transporting all the goods from the suppliers to the consumers. Thus, the OT problem can be formulated as follows:

$$\min_{\Gamma_{ij} \in \Gamma} \sum\nolimits_{i,j}^{m,n} \Gamma_{ij} c_{ij}, \tag{1}$$

where $\Gamma_{ij} \geq 0$. The constraints to be satisfied are: the i-th supplier holds $x_i = \sum_{j=1}^{n} \Gamma_{ij}$ units of goods, and the j-th consumer needs $y_j = \sum_{i=1}^{m} \Gamma_{ij}$ units goods. Meanwhile, the total amount of goods held by all suppliers are equal to the amount needed by all consumers, *i.e.*, $\sum_{i=1}^{m} x_i = \sum_{j=1}^{n} y_j$. To efficiently tackle this problem, we adopt the Sinkhorn Iteration method [12]. The details can be found in the Supplementary Materials.

3.3. Pseudo-mask Generation by OT

Given an input image $I^{H \times W \times 3}$, supposing there are m gt point labels and n pixel samples (*i.e.*, $n = H \times W$), we view each gt point label as a supplier who holds k pixel samples (*i.e.*, $x_i = k, i = 1, 2, ..., m$). Each pixel of I is regarded as a consumer who needs one gt point label (*i.e.*, $y_j = 1, j = 1, 2, ..., n$). Given the defined cost c_{ij} to transport one unit from the $i\text{-}th$ gt point label to the j-th pixel, the global OT plan $\Gamma \in \mathbb{R}^{m \times n}$ can be obtained by solving the OT problem via the Sinkhorn-Knopp Iteration [12]. Once Γ is obtained, the pseudo-mask label generation can be decoded by assigning the pixel samples to the suppliers who transport point gt labels to them with the minimal transportation costs.

The pseudo-mask generation consists of task-oriented map generation, transportation cost definition and centroid-based unit number calculation, which are introduced in details in the following subsections. The completed procedure is summarized in Algorithm 1.

3.3.1 Task-oriented Map Generation

The task-oriented map includes the category-wise semantic map P^s and instance-wise boundary map P^b. The former measures the semantic logit differences among the various categories. The latter discriminates the different thing-based targets under the same class from the accurate instance-level boundary. Based on these maps, the distance of the adjacent pixels can be calculated to obtain each pixel-to-gt cost c_{ij}.

Category-wise Semantic Map. An input image for panoptic segmentation task is composed of the stuff-based and thing-based targets. The semantic parsing is important to obtain category-wise logits. As shown in Fig. 3, we adopt the transformer decoder layers [29] to construct the semantic decoder with a set of semantic query tokens, which is one-to-one match to the semantic categories. The semantic logits P^s with N_c classes can be generated by multiplying the mask scores and the class probabilities together as in [15]. The supervision information for category-wise semantic logits P^s with the weak point labels is introduced in Sec. 3.4.1 in detail.

Instance-wise Boundary Map. To discriminate the instances for thing-based targets, especially for the instances

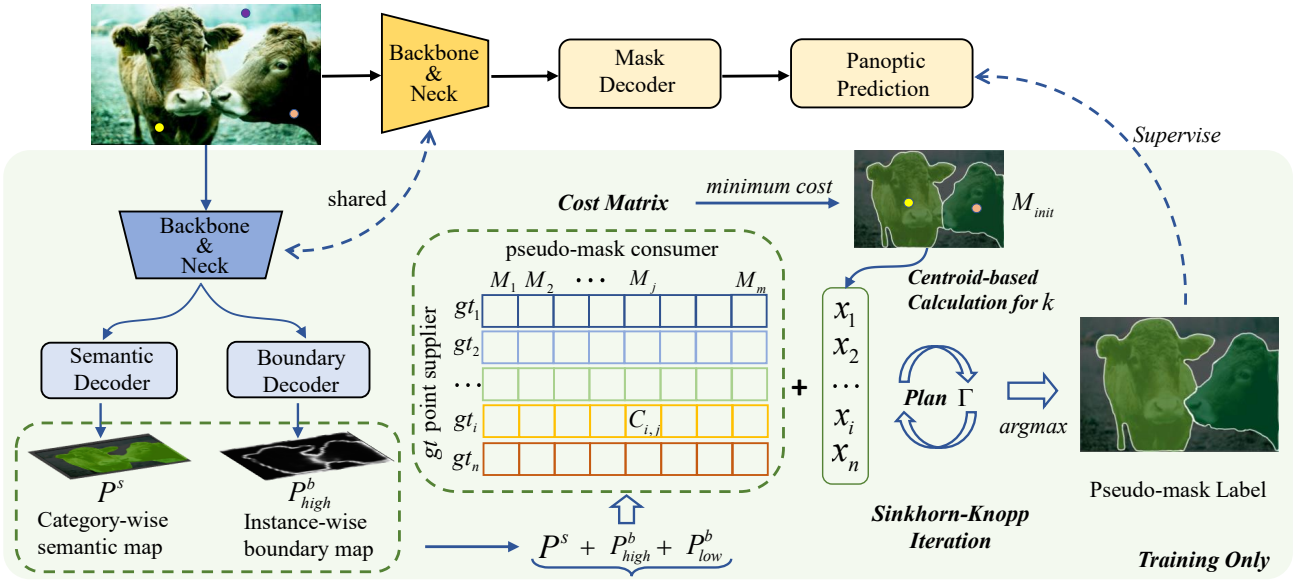

Figure 3: Overview of Point2Mask. It consists of two branches, one branch for mask pseudo-label generation, and another for panoptic segmentation based on the generated pseudo-labels. The mask pseudo-label generation is formulated as the OT problem, where the cost matrix is defined based on the task-oriented maps. The k unit number is calculated by the centriod-based scheme. The global optimal transportation plan Γ can be solved by the Sinkhorn-Knopp Iteration to obtain the accurate pseudo-mask labels. Only panoptic segmentation branch is kept for inference.

with the same category, we introduce the instance-wise boundary map P^b for each target.

To generate the pure boundary, we suggest the high-level boundary P^b_{high} that is learnt by the boundary decoder. In specific, we firstly sum the multi-level feature tokens from the Transformer-based neck in 2D spatial feature. Then, two 1×1 convolution layers interleaved by a ReLU activation are employed. The one-channel boundary map P^b_{high} is obtained via the `sigmoid` function. For high-level boundary learning objective, we design an effective boundary loss function and explain it with details in Sec. 3.4.1.

Besides, we employ the Structured Edge (SE) detection method [13] based on the original input image to capture the low-level contour P^b_{low}, which takes advantage of the inherent structure in edge patches to focus on the sparse object-level boundary map.

3.3.2 Transportation Cost

Based on the obtained task-oriented maps, the transportation cost can be calculated.

In our method, each map can be represented as an 8-connected planar graph $G(V, E)$, where each pixel is adjacent to eight neighbors. The vertex set V consists of all pixels of the map, and the edge set E is made of the edges between two adjacent vertices. Let the vertex l and vertex k be adjacent on the graph. Based on the P^s and P^b maps, the corresponding distance function $d^s_{k,l}$ and $d^b_{k,l}$ can be defined as follows:

$$\begin{aligned} d^s_{k,l} &= |P^s(k) - P^s(l)|\,, \\ d^b_{k,l} &= \max\{P^b(k), P^b(l)\}, \end{aligned} \tag{2}$$

where $P(l)$, $P(k)$ are the map values of vertex l and vertex k, respectively. Once the edge length is obtained from the P^s and P^b maps, we define the transportation cost $c_{i,j}$ from the i-th pixel to the j-th gt point label as the sum of the lengths of their connected edges along the shortest path $\mathbb{P}$:

$$c_{i,j} = \sum_{(k,l)\in\mathbb{P}_{i,j}} (d^s_{k,l} + \beta d^b_{k,l}), \tag{3}$$

where β is the balanced weight. The shortest path $\mathbb{P}$ is implemented by the classical `Dijkstra` algorithm like [15].

3.3.3 Centroid-based Unit Number Calculation

Each gt point label $\mathcal{P}_i$ is regarded as the supplier in our proposed OT problem, which holds $x_i = k$ pixels of pseudo mask label M. To set the accurate number of k, we introduce the centroid-based unit number calculation scheme that can be divided into two steps, as shown in Fig. 4.

Firstly, we obtain the pair-wise cost values along the shortest path $\mathbb{P}$ for each undetermined pixel to each gt point label $\mathcal{P}_i$. The initial gt point label assignment for each pixel

Algorithm 1 Optimal Transport for Pseudo-mask Generation

Input:

$I^{H\times W\times 3}$ is an input image.

$M^{H\times W\times 1}$ is the pseudo-mask label with ZerosInit.

$\mathcal{P}$ is a set of gt point labels.

T is the iteration number in Sinkhorn-Knopp Iter.

Output:

M is the assigned pseudo-mask label.

$m \leftarrow |\mathcal{P}|, n \leftarrow |M|$

$P^s, P^b_{high}, P^b_{low} \leftarrow$ Forward$(I, \mathcal{P})$

Compute pairwise pixel-to-gt cost c_{ij}.

$x_i(i = 1, 2, ..., m) \leftarrow$ Centriod-based k calculation

$y_j(j = 1, 2, ..., n) \leftarrow \mathbb{1}$ ▷ Init y with ones

$u^0, v^0 \leftarrow \mathbb{1}$ ▷ Init u and v with ones

for $t = 0$ **to** T **do**:

  $u^{t+1}, v^{t+1} \leftarrow$ SinkhornIter(c, u^t, v^t, x, y)

Compute optimal plan Γ.

Compute pseudo-mask label: $M = \text{argmax}(\Gamma)$.

return M

can be achieved with its minimum cost among all gt labels in the whole image. Note that the gt points are randomly labeled on each target in the image, which can be located at any position of the target to be segmented, such as the corner or the edge. This cannot reflect the typical and accurate characteristics, especially for the border pixels between thing-based instances belonging to the same category.

Based on the initial gt point label assignment, the initial mask label for each target can be obtained. We then calculate the corresponding centroid $\mathcal{C}_i$ of initial mask label as the substitution of gt point label $\mathcal{P}_i$ for each target. The pairwise cost c_{ij} for each pixel and $\mathcal{C}_i$ can be re-calculated along the corresponding shortest path. The k unit number (x_i) is computed by counting the ones in N_{ij} with the minimum cost values to each centriod $\mathcal{C}$, which can be formulated as follows:

$$x_i = \sum_j^n N_{ij}, \quad N_{ij} = \begin{cases} 1, & \underset{i}{\text{argmin}}\, c_{ij} = i, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The iterated calculation scheme can obtain a more accurate unit number k, and we leave the detailed performance analysis in Sec. 4.4 to examine the effectiveness of the proposed scheme.

3.4. Learning and Inference

3.4.1 Weakly Supervised Learning

In this section, we introduce the objective for category-wise semantic map P^s and instance-wise boundary map P^b in a weakly-supervised manner with only a single point label.

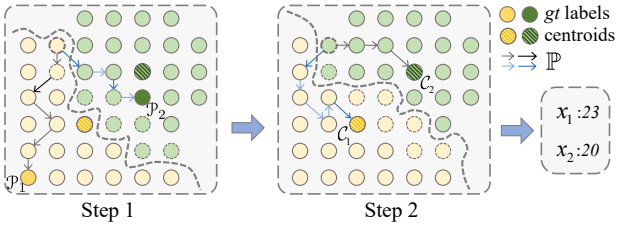

Figure 4: The process of centroid-based k calculation with two targets in an image. **Step 1:** The initial assignment (*i.e.*, the pixels with yellow and green color divided by the middle curve line of dashes) with the minimal cost can be achieved based on the gt point labels $\mathcal{P}_1$ and $\mathcal{P}_2$. **Step2:** The centroids $\mathcal{C}_1$ and $\mathcal{C}_2$ of each initially assigned mask are the substitutions of gt points, and the minimal cost can be re-calculated to achieve the refined assignment and determine the accurate unit number k for each target.

Semantic Map Learning. Like the weakly-supervised semantic methods [30, 42], we adopt the partial cross-entropy loss $\mathcal{L}_{partial}$, which is able to make full use of the available gt point labels to achieve region supervised learning and generate sparse semantic map.

To obtain the accurate semantic logits for the unlabeled regions, we further take advantage of both local LAB affinity and long-range RGB affinity based on the input image. Local LAB affinity explores the color similarity in LAB color space with the local kernel, which is employed as the loss term $\mathcal{L}^{LAB}_{sem}$ as in [43]. Long-range RGB affinity absorbs the pixel similarity in RGB space, which is implemented by the minimum spanning tree. As in [30], it is utilized as the loss term $\mathcal{L}^{RGB}_{sem}$. The objective for semantic map learning is denoted as:

$$\mathcal{L}_{sem} = \mathcal{L}_{partial} + \alpha_1 \mathcal{L}^{LAB}_{sem} + \alpha_2 \mathcal{L}^{RGB}_{sem}. \tag{5}$$

Please refer to the Supplementary Materials for the detailed formulation of these loss terms.

High-level Boundary Map Learning. To encourage the boundary decoder to predict the high-level instance-wise boundary map P^b_{high}, we suggest an effective loss function $\mathcal{L}_{bou}$ for panoptic segmentation task. In terms of the existence of a boundary between two adjacent pixels, we assume that their affinity is small as in [1]. Hence, we introduce the high-level affinity $\mathcal{A}$ representation. For each pixel p_k on P^b_{high}, p_l is one of its eight neighbors $\mathcal{N}_8$. The $\mathcal{A}_{kl}$ can be represented as follows:

$$\mathcal{A}_{kl} = 1 - \max P^b_{high}(p_k, p_l). \tag{6}$$

Then, we make full use of the mask affinity equivalence among the neighbor pixels based on the generated pseudo-

mask M. The loss function $\mathcal{L}_{bou}$ can be defined as:

$$\mathcal{L}_{bou} = -\sum_{(k,l)\in M^{+}_{thing}} \frac{\log \mathcal{A}_{kl}}{2\left|M^{+}_{thing}\right|} - \sum_{(k,l)\in M^{+}_{stuff}} \frac{\log \mathcal{A}_{kl}}{2\left|M^{+}_{stuff}\right|} - \sum_{(k,l)\in M^{-}} \frac{\log(1-\mathcal{A}_{kl})}{|M^{-}|}, \tag{7}$$

where M^{+}_{thing} denotes that the pair of adjacent pixels p_k and p_l are inside the same thing-based pseudo mask. Similarly, M^{+}_{stuff} represents that p_k and p_l are inside the same stuff-based pseudo mask. Instead, M^{-} denotes that a pair of pixels are with different pseudo-mask labels. Driven by the $\mathcal{L}_{bou}$ term, we can learn the accurate high-level boundary. The Supplementary Materials show some visual examples for better illustration.

3.4.2 Training and Inference

Loss Function. Once the pseudo-masks are obtained, the panoptic segmentation sub-model is trained with these generated labels in a fully supervised manner. We adopt Panoptic SegFormer [29] as the panoptic sub-network. The fully-supervised loss terms consist of the focal loss for classification prediction, the localization loss for box localization, and the dice loss on mask decoder for final panoptic segmentation, respectively. For simplicity, we denote these losses to train the panoptic segmentation model as $\mathcal{L}_{full}$. The total loss $\mathcal{L}_{total}$ can be formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{full} + \mathcal{L}_{sem} + \mathcal{L}_{bou}. \tag{8}$$

Inference. For the inference process of Point2Mask, only the panoptic segmentation model is maintained after training, which is the same as the original Panoptic SegFormer model [28]. The process of pseudo-mask generation with OT incurs about 25% extra computational load in training, but it is totally cost-free during inference.

4. Experiments

To evaluate our proposed approach, we conduct experiments on Pascal VOC [14] and COCO [31]. *Only a single point label per target is used to train our method*, which is randomly sampled with the uniform distribution from the original pixel-wise mask annotations.

4.1. Datasets

Pascal VOC [14]. Pascal VOC consists of 20 "thing" and 1 "stuff" categories. It contains 10,582 images for model training and 1,449 validation images for evaluation [17].
COCO [31]. COCO has 80 "thing" and 53 "stuff" categories, which is a challenging benchmark. Our models are trained on `train2017` (115K images), and evaluated on `val2017` (5K images).

4.2. Implementation Details

The models are trained with the AdamW optimizer [33]. We make use of the `mmdetection` toolbox [4] and follow the commonly used training settings on each dataset. ResNet [18] and Swin-Transformer [32] are employed as the backbones, which are pre-trained on ImageNet [36]. On Pascal VOC, the initial learning rate is set to 10^{-4}, and the weight decay is 0.1 with eight images per mini-batch. The models are trained with $2\times$ schedule at 24 epochs. On COCO, the initial learning rate is set to 2×10^{-4}, which is reduced by a factor of 10 at the 8-th epoch and 12-th epoch with 16 images per mini-batch. The models are trained with 15 epochs. The iteration number in Sinkhorn Iteration for solving the defined OT problem is set to 80. β is 0.1 in Eq. 3, and $\alpha_1 = \alpha_2 = 3.0$ in Eq. 5 in our implementation. As in [28], the number of query tokens for fully panoptic segmentation sub-model is set to 300. The manifold projector proposed in [15] is employed to better stand for the instance-wise representation based on our baseline model. Unless specified, our centroid-based unit number calculation scheme is not iterated in the main experiments. We report the standard evaluation metrics [21] of panoptic segmentation task, including panoptic quality (PQ), segmentation quality (SQ) and recognition quality (RQ).

4.3. Main Results

We compare our proposed Point2Mask method against state-of-the-art weakly supervised panoptic segmentation approaches. Moreover, the results of representative fully mask-supervised methods are reported for reference.

Results on Pascal VOC. Table 1 reports the comparison results on Pascal VOC `val`. It can be clearly seen that Point2Mask with the ResNet-50 backbone outperforms the recent single point-supervised method PSPS [15] by absolute 4.0% PQ (from 49.8% to 53.8%). The performance improvement mainly stems from the thing-based objects, from 47.8% PQ^{th} to 51.9% PQ^{th} (+4.1% PQ^{th}), in contrast to the improvements on PQ^{st} (89.5% *vs.* 90.3%). It demonstrates the effectiveness of our presented pseudo-mask generation scheme by OT for thing-based instances. Our approach even outperforms Panoptic FCN [28] with 10 point labels by 5.8% PQ (53.8% *vs.* 48.0%). Moreover, our proposed method obtains 61.0% PQ with Swin-L [32] backbone, which achieves comparable results against the fully supervised methods. When the point-label COCO dataset is used for model pre-training, we achieve significant performance improvements, such as from 53.8% PQ to 60.7% PQ under the ResNet-50 backbone. With the Swin-L backbone, Point2Mask obtains 64.2% PQ, surpassing the fully supervised method [25] by 1.1% PQ.

Results on COCO. Table 2 gives the evaluation results comparing to the state-of-the-art (SOTA) methods on COCO. Our proposed Point2Mask method achieves 32.4%

Method	Backbone	Supervision	VOC 2012			VOC 2012 *with* COCO		
			PQ	PQ^{th}	PQ^{st}	PQ	PQ^{th}	PQ^{st}
Li *et al*. [25]	ResNet-101	$\mathcal{M}$	62.7	-	-	63.1	-	-
Panoptic FPN [21]	ResNet-50	$\mathcal{M}$	65.7	64.5	90.8	-	-	-
Panoptic FCN [28]	ResNet-50	$\mathcal{M}$	67.9	66.6	**92.9**	**73.1**	**72.1**	**93.8**
Panoptic SegFormer [29]	ResNet-50	$\mathcal{M}$	**67.9**	**66.6**	92.7	-	-	-
Li *et al.* [25]	ResNet-101	$\mathcal{B}+\mathcal{I}$	59.0	-	-	59.5	-	-
JTSM [38]	ResNet-18-WS [39]	$\mathcal{I}$	39.0	37.1	77.7	-	-	-
PSPS [15]	ResNet-50	$\mathcal{P}$	49.8	47.8	89.5	-	-	-
Panoptic FCN [28]	ResNet-50	$\mathcal{P}_{10}$	48.0	46.2	85.2	52.4	50.8	86.0
Point2Mask	ResNet-50	$\mathcal{P}$	53.8	51.9	90.5	60.7	59.1	91.8
Point2Mask	ResNet-101	$\mathcal{P}$	54.8	53.0	90.4	63.2	61.8	92.3
Point2Mask	Swin-L	$\mathcal{P}$	**61.0**	**59.4**	**93.0**	**64.2**	**62.7**	**93.2**

Table 1: Performance comparisons on Pascal VOC2012 `val`. $\mathcal{M}$ denotes the pixel-wise mask annotations. $\mathcal{P}$ and $\mathcal{P}_{10}$ are point-level supervision with 1 and 10 points per target, respectively. $\mathcal{I}$ and $\mathcal{B}$ are the image-level and box-level supervisions (the same below). Besides, VOC 2012 *with* COCO represents training and validation on VOC 2012 dataset with COCO pre-trained model.

Method	Backbone	Supervision	PQ	PQ^{th}	PQ^{st}	SQ	RQ
AdaptIS [40]	ResNet-50	$\mathcal{M}$	35.9	40.3	29.3	-	-
Panoptic FPN [21]	ResNet-50	$\mathcal{M}$	39.4	45.9	29.6	77.8	48.3
Panoptic-DeepLab [7]	Xception-71 [11]	$\mathcal{M}$	39.7	43.9	33.2	-	-
Panoptic FCN [28]	ResNet-50	$\mathcal{M}$	43.6	49.3	35.0	**80.6**	**52.6**
Panoptic SegFormer [29]	ResNet-50	$\mathcal{M}$	48.0	52.3	41.5	-	-
Mask2Former [8]	ResNet-50	$\mathcal{M}$	**51.9**	**57.7**	**43.0**	-	-
JTSM [38]	ResNet-18-WS	$\mathcal{I}$	5.3	8.4	0.7	30.8	7.8
PSPS [15]	ResNet-50	$\mathcal{P}$	29.3	29.3	29.4	-	-
Panoptic FCN [28]	ResNet-50	$\mathcal{P}_{10}$	31.2	35.7	24.3	-	-
Point2Mask	ResNet-50	$\mathcal{P}$	32.4	32.6	32.2	75.1	41.5
Point2Mask	ResNet-101	$\mathcal{P}$	34.0	34.3	33.5	75.1	43.5
Point2Mask	Swin-L	$\mathcal{P}$	**37.0**	**37.0**	**36.9**	**75.8**	**47.2**

Table 2: Panoptic segmentation results on COCO `val2017`. Weakly and fully supervised methods are compared.

PQ with single point supervision when ResNet-50 is employed as the backbone. It outperforms the previous SOTA method PSPS [15] by 3.1% PQ, 3.3% PQ^{th} and 2.8% PQ^{st} under the same setting. Compared with Panoptic FCN [28] with 10 point labels, our approach surpasses it by 1.2% PQ (32.4% *vs.* 31.2%). With Swin-L as the backbone, Point2Mask achieves 37.0% PQ performance, which is comparable with some fully mask-supervised methods, including AdaptIS [40], Panoptic FPN [21] and Panoptic-DeepLab [7] with ResNet-50 backbone.

4.4. Ablation Studies

We analyze the design of each component in Point2Mask on Pascal VOC dataset.

Different Task-oriented Maps. We employ the category-wise semantic map P^s, low-level and high-level boundary map P^b_{low}, P^b_{high} to calculate the cost for optimal transport. Table 3 shows the evaluation results with different task-oriented maps. Our method achieves 50.6% PQ using the P^s map only, which focuses on the semantic logit differences among the categories. When P^b_{low} and P^b_{high} are employed separately, our method achieves 51.1% PQ and 53.4% PQ, respectively. More specifically, P^b_{high} brings +2.9% PQ gains driven by the designed boundary loss function $\mathcal{L}_{bou}$. When all maps are adopted, Point2Mask achieves the best performance of 53.8% PQ.

Semantic Map Learning. Single point-supervised semantic parsing is the bedrock to obtain the panoptic segmentation results in our Point2Mask. As shown in Table 4, when both local LAB loss $\mathcal{L}^{LAB}_{sem}$ and long-range RGB loss $\mathcal{L}^{RGB}_{sem}$ are adopted for the semantic map learning, the best 69.5% mIoU and 53.8% PQ are obtained comparing to each individual loss term.

Different Unit Number Calculation Schemes. We explore three different schemes to calculate the unit number k for gt supplier, including "Equal Division", "Nearest gt

P^s	P^b_{low}	P^b_{high}	PQ	PQth	PQst
✓			50.6	48.7	90.1
✓	✓		51.1	49.1	90.3
✓		✓	53.4	51.6	90.3
✓	✓	✓	**53.8**	**51.9**	**90.5**

Table 3: The impact of different task-oriented maps to calculate the pixel-to-gt point label cost c_{ij} in OT.

$\mathcal{L}_{partial}$	$\mathcal{L}^{LAB}_{sem}$	$\mathcal{L}^{RGB}_{sem}$	mIoU	PQ	PQth	PQst
✓			61.6	40.4	38.1	86.1
✓	✓		69.0	51.2	49.3	90.0
✓		✓	68.0	49.5	47.5	89.3
✓	✓	✓	**69.5**	**53.8**	**51.9**	**90.5**

Table 4: Comparison of different weakly-supervised loss terms for category-wise semantic map learning.

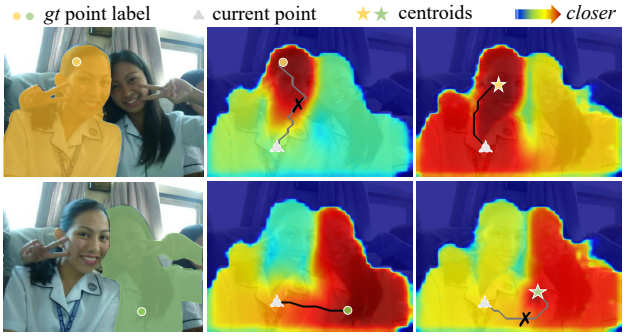

Figure 5: Visual comparisons on distance heatmap with different calculation schemes of k. (a) shows the gt point label and pixel-wise mask label. (b) indicates the heatmap based on the Nearest gt Point scheme. (c) is the heatmap based on our proposed Nearest Centroid scheme. The corresponding shortest paths are shown for better illustration.

Point" and "Nearest Centroid". The Equal Division treats the mean value as k for each gt point supplier from all pixels. The Nearest gt Point indicates that the total number of pixels are with the nearest distances measured by the cost for each gt point. For simplicity, we denote the presented centroid-based unit number calculation scheme in Sec. 3.3.3 as the Nearest Centroid. Table 5 reports the comparison results. Our Nearest Centroid scheme obtains the best performance with 53.8% PQ, which outperforms Equal Division and Nearest gt Point by 1.4% PQ and 1.0% PQ, respectively. Furthermore, we report the visual comparisons on distance heatmap, as shown in Fig. 5. It can be clearly seen that the proposed Nearest Centroid scheme obtains the accurate unit number k for each gt point supplier.

In addition, as shown in Table 6, the Nearest Centroid scheme with more iterations (8 iterations) can bring a performance gain of +0.48% PQ. With 10 iterations, the model achieves the saturated performance with 54.07% PQ.

Different Pseudo-mask Generation Methods. To examine the effectiveness of our proposed OT-based scheme, we study the different methods on pseudo-mask generation in Point2Mask. Based on the presented cost on the task-oriented maps, we compared OT with the direct minimum cost (MC) method. Similar to [15], MC assigns the gt point label to each pixel with its corresponding minimum cost individually. Table 7 shows the comparison results. Point2Mask with our proposed OT method outperforms the MC scheme by +2.3% PQ. Specifically, the performance gains mainly stem from the thing-based targets (+2.3% PQth *vs.* +0.1% PQst). This is because it takes consideration of the global optimization in dealing with the ambiguous locations, like the border pixels between different thing-based targets with the same category.

Scheme	PQ	PQth	PQst
Equal Division	52.4	50.5	90.2
Nearest gt Point	52.8	50.9	90.1
Nearest Centriod	**53.8**	**51.9**	**90.5**

Table 5: Performance with different calculation schemes of k for our defined OT problem in Point2Mask.

Iterations	1	2	4	8	10
PQ	53.76	53.80	53.91	**54.24**	54.07

Table 6: Performance with various iterations in centroid updating of the Nearest Centroid scheme.

Method	PQ	PQth	PQst
Minimum Cost	51.9	50.1	90.2
Optimal Transport	**54.2**(↑2.3)	**52.4**(↑2.3)	**90.3**(↑0.1)

Table 7: Comparisons between Minimum Cost (MC) and Optimal Transport (OT) based on the defined cost for pseudo-mask label generation.

5. Conclusion

An effective single point-supervised panoptic segmentation approach, namely Point2Mask, was presented. The accurate pseudo-mask was obtained by finding the optimal transport plan at a globally minimal transportation cost, which was defined according to the task-oriented maps. Moreover, an effective centroid-based scheme was introduced to obtain the accurate unit number for each gt point supplier. Extensive experiments were conducted on Pascal VOC and COCO benchmarks, validating the leading performance of the proposed Point2Mask over the previous state-of-the-arts on point-supervised panoptic segmentation.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grants (61831015). Corresponding author is Jianke Zhu.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2209–2218, 2019.

[2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *Proc. Eur. Conf. Comp. Vis.*, pages 549–565. Springer, 2016.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. Eur. Conf. Comp. Vis.*, pages 213–229. Springer, 2020.

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

[5] Pengfei Chen, Xuehui Yu, Xumeng Han, Najmul Hassan, Kai Wang, Jiachen Li, Jian Zhao, Humphrey Shi, Zhenjun Han, and Qixiang Ye. Point-to-box network for accurate object detection via single point supervision. In *Proc. Eur. Conf. Comp. Vis.*, pages 51–67. Springer, 2022.

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Proc. Eur. Conf. Comp. Vis.*, pages 104–120. Springer, 2020.

[7] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 12475–12485, 2020.

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1290–1299, 2022.

[9] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2617–2626, 2022.

[10] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Proc. Advances in Neural Inf. Process. Syst.*, volume 34, pages 17864–17875, 2021.

[11] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1251–1258, 2017.

[12] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proc. Advances in Neural Inf. Process. Syst.*, volume 26, 2013.

[13] Piotr Dollár and C Lawrence Zitnick. Structured forests for fast edge detection. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1841–1848, 2013.

[14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, 2010.

[15] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Pointly-supervised panoptic segmentation. In *Proc. Eur. Conf. Comp. Vis.*, pages 319–336. Springer, 2022.

[16] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 303–312, 2021.

[17] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 991–998, 2011.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 770–778, 2016.

[19] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella X Yu. Universal weakly supervised segmentation by pixel-to-segment contrastive learning. In *Proc. Int. Conf. Learning Represent.*, 2021.

[20] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 6399–6408, 2019.

[21] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 6399–6408, 2019.

[22] 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 *Proc. IEEE Int. Conf. Comp. Vis.*, pages 3406–3416, 2021.

[23] Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, Lionel M Ni, Heung-Yeung Shum, et al. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2023.

[24] Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 13936–13944, 2020.

[25] Qizhu Li, Anurag Arnab, and Philip HS Torr. Weakly-and semi-supervised panoptic segmentation. In *Proc. Eur. Conf. Comp. Vis.*, pages 102–118, 2018.

[26] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Xian-Sheng Hua, and Lei Zhang. Box-supervised instance segmentation with level set evolution. In *Proc. Eur. Conf. Comp. Vis.*, pages 1–18. Springer, 2022.

[27] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Risheng Yu, Xiansheng Hua, and Lei Zhang. Box2mask: Box-supervised instance segmentation via level-set evolution. *arXiv preprint arXiv:2212.01579*, 2022.

[28] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Yukang Chen, Lu Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation with point-based supervision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[29] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Ping Luo, and Tong Lu.

Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1280–1289, 2022.

[30] Zhiyuan Liang, Tiancai Wang, Xiangyu Zhang, Jian Sun, and Jianbing Shen. Tree energy loss: Towards sparsely annotated semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 16907–16916, 2022.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. Eur. Conf. Comp. Vis.*, pages 740–755. Springer, 2014.

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 10012–10022, 2021.

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. Int. Conf. Learning Represent.*, 2019.

[34] Svetlozar T Rachev. The monge–kantorovich mass transference problem and its stochastic applications. *Theory of Probability & Its Applications*, 29(4):647–676, 1985.

[35] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A metric for distributions with applications to image databases. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 59–66. IEEE, 1998.

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3):211–252, 2015.

[37] Wei Shen, Zelin Peng, Xuehui Wang, Huayu Wang, Jiazhong Cen, Dongsheng Jiang, Lingxi Xie, Xiaokang Yang, and Q Tian. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[38] Yunhang Shen, Liujuan Cao, Zhiwei Chen, Feihong Lian, Baochang Zhang, Chi Su, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Toward joint thing-and-stuff mining for weakly supervised panoptic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 16694–16705, 2021.

[39] Yunhang Shen, Rongrong Ji, Yan Wang, Zhiwei Chen, Feng Zheng, Feiyue Huang, and Yunsheng Wu. Enabling deep residual networks for weakly supervised object detection. In *Proc. Eur. Conf. Comp. Vis.*, pages 118–136. Springer, 2020.

[40] Konstantin Sofiiuk, Olga Barinova, and Anton Konushin. Adaptis: Adaptive instance selection network. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 7355–7363, 2019.

[41] Chufeng Tang, Lingxi Xie, Gang Zhang, Xiaopeng Zhang, Qi Tian, and Xiaolin Hu. Active pointly-supervised instance segmentation. In *Proc. Eur. Conf. Comp. Vis.*, pages 606–623. Springer, 2022.

[42] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1818–1827, 2018.

[43] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5443–5452, 2021.

[44] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5463–5474, 2021.

[45] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 8818–8826, 2019.

[46] Xue Yang, Gefan Zhang, Wentong Li, Xuehui Wang, Yue Zhou, and Junchi Yan. H2rbox: Horizonal box annotation is all you need for oriented object detection. 2023.

[47] Qihang Yu, Huiyu Wang, Dahun Kim, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Cmt-deeplab: Clustering mask transformers for panoptic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2560–2570, 2022.

[48] Xuehui Yu, Pengfei Chen, Di Wu, Najmul Hassan, Guorong Li, Junchi Yan, Humphrey Shi, Qixiang Ye, and Zhenjun Han. Object localization under single coarse point supervision. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4868–4877, 2022.

[49] Yi Yu, Xue Yang, Qingyun Li, Yue Zhou, Gefan Zhang, Junchi Yan, and Feipeng Da. H2rbox-v2: Boosting hbox-supervised oriented object detection via symmetric learning. *arXiv preprint arXiv:2304.04403*, 2023.

[50] Bingfeng Zhang, Jimin Xiao, Jianbo Jiao, Yunchao Wei, and Yao Zhao. Affinity attention graph neural network for weakly supervised semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[51] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. In *Proc. Advances in Neural Inf. Process. Syst.*, volume 34, pages 10326–10338, 2021.