Pointly-Supervised Panoptic Segmentation

Junsong Fan[1,3], Zhaoxiang Zhang*[1,2,3], and Tieniu Tan[1,2]

[1] Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Centre for Artificial Intelligence and Robotics, HKISI_CAS, HongKong, China
{fanjunsong2016@,zhaoxiang.zhang@,tnt@nlpr.}ia.ac.cn

Abstract. In this paper, we propose a new approach to applying point-level annotations for weakly-supervised panoptic segmentation. Instead of the dense pixel-level labels used by fully supervised methods, point-level labels only provide a single point for each target as supervision, significantly reducing the annotation burden. We formulate the problem in an end-to-end framework by simultaneously generating panoptic pseudo-masks from point-level labels and learning from them. To tackle the core challenge, i.e., panoptic pseudo-mask generation, we propose a principled approach to parsing pixels by minimizing pixel-to-point traversing costs, which model semantic similarity, low-level texture cues, and high-level manifold knowledge to discriminate panoptic targets. We conduct experiments on the Pascal VOC and the MS COCO datasets to demonstrate the approach's effectiveness and show state-of-the-art performance in the weakly-supervised panoptic segmentation problem. Codes are available at https://github.com/BraveGroup/PSPS.git.

Keywords: weakly-supervised learning, panoptic segmentation

1 Introduction

Panoptic segmentation [23] aims at fully parsing all the pixels into nonoverlapping masks for both thing instances and stuff classes. It combines the semantic segmentation and the instance segmentation tasks simultaneously. Classical deep learning approaches require precise dense pixel-level labels to solve this problem. However, acquiring exact pixel- and instance-level annotations on large-scale datasets is very time-consuming, hindering the popularization and generalization of the approaches in new practical applications.

To alleviate the annotation burden for segmentation models, researchers recently proposed weakly-supervised learning (WSL) [4,52,51], which focuses on leveraging coarse labels to train dense pixel-level segmentation tasks. Typically, the weak supervision includes image-level [14,16,15], point-level [2,38], scribble-level [47,31], and bounding box-level labels [9], etc. These approaches tackle either semantic segmentation [36], instance segmentation [1,21], or panoptic segmentation [41,27] tasks. Among them, the weakly-supervised panoptic segmentation (WSPS) problem is the most challenging since it requires both semantic

and instance discrimination with only weak supervision. As a result, the WSPS got less attention in previous works, and its performance is far from satisfactory. The seminal work by Li et al. [27] manages to address the WSPS problem using bounding-box level labels. Later, JTSM [41] proposes to apply only image-level labels for the WSPS problem. Recently, PanopticFCN [29] tackles this problem by connecting multiple point labels into polygons. The performances of these approaches differ significantly with the different weak annotations.

In this paper, we propose a new WSPS paradigm to use only a single point for each target as the supervision, as illustrated in Fig. 1. Recall that the core of weakly-supervised segmentation is to release the annotation burden while still obtaining decent performance. In other words, balance the cost of annotation and the model performance. We are motivated to use the point-level labels because, on the one hand, the annotation time of point-level labels is only marginally above the image-level labels [2], saving much cost compared with the box-level or polygon labels. On the other hand, point labels can provide minimum spatial information to localize and discriminate different panoptic targets for the segmentation models.

A natural idea to estimate panoptic masks from point-level labels is to assign each pixel in the image to one of the points according to some principles. To this end, we propose tackling this problem by minimizing the pixel-to-point traversing cost, measured by the neighboring pixel affinities. There are two basic requirements to correctly assign pixels to point labels: semantic class matching and instance discrimination. The former ensures that the pixels are parsed with the correct class labels, and the latter is responsible for distinguishing different instances in the thing classes. Therefore, we consider three criteria to model the affinities: semantic similarity, low-level image cues, and high-level instance discrimination knowledge. Using these criteria, we model the pixel-to-point traversing costs and solve the assignment problem by finding the shortest path.

We base our approach on the transformer models [46,11,39], which have recently shown impressive results in computer vision tasks [8,30,44,3,56]. Specifically, our approach contains a group of semantic query tokens to parse semantic segmentation results and a group of panoptic query tokens responsible for the panoptic segmentation task [30]. In addition to the regular panoptic segmentation model, our approach contains a label generation model, which produces dense panoptic pseudo-masks depending on the point-level labels and the criteria above. The whole approach is end-to-end. After training, only the panoptic segmentation branch is kept for testing. Thus, it does not incur additional computation or memory overhead for usage. We conduct thorough experiments to analyze the proposed approach and the properties of the point-level labels. Meanwhile, we demonstrate new cutting-edge performance with the WSPS problem on the Pascal VOC [13] and the MS COCO [32] datasets.

In summary, the main contributions of this work are:

- We propose a new paradigm for the WSPS problem, which utilizes a single point for each target as supervision for training.

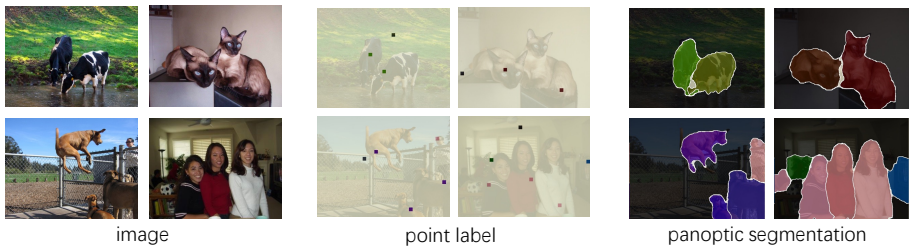

Fig. 1. Illustration of the proposed pointly-supervised panoptic segmentation. From left to right: input images, point labels as supervision, and panoptic segmentation predictions. The point labels provide a single point annotation for each target, including both thing instances and stuff classes, which are used at training time only. Please see Sec. 3 for details. Best viewed in color.

- A novel approach to estimating dense panoptic pseudo-masks by minimizing the pixel-to-point traversing distance is proposed.
- We implement the approach in an end-to-end framework with transformers, conduct analytical experiments to study the model and the point-level labels, and demonstrate state-of-the-art performance on the Pascal VOC and the MS COCO datasets.

2 Related Works

2.1 Panoptic Segmentation

The panoptic segmentation [23] task simultaneously incorporates semantic segmentation and instance segmentation, where each pixel is uniquely assigned to one of the stuff classes or one of the thing instances. This problem can be tackled by combining the semantic and instance segmentation results in a post-processing manner [23]. Later works such as JSIS [10] adopt a unified network combining a semantic segmentation branch and an instance segmentation branch. After that, many approaches have been proposed for improvement by using feature pyramids [22], automatic architecture searching [49], and unifying the pipeline [28], etc.

Recently, transformer-based approaches have shown impressive results across the NLP [46,11,39] and the CV [3,12,56,34] applications. The seminal work DETR [3] provides a clear and elegant solution for object detection and segmentation. The following work DeformableDETR [56] improves it by using the deformable transformers to reduce the computation burden and accelerate the convergence. K-Net [54] adopts an iterative refinement procedure to enhance the attention masks gradually. MaskFormer [8] proposes to separate the mask prediction and the classification process. Panoptic SegFormer [30] embraces a similar idea and adopts an auxiliary localization target to ease the model training. Our panoptic segmentation approach is based on these works, and we focus on alleviating the annotation burden by exploiting point-level annotations.

2.2 Weakly-Supervised Segmentation

Weakly supervised segmentation [4,52] aims to alleviate the annotation burden for segmentation tasks by using weak labels for training. According to the type of tasks, it concerns semantic segmentation [16,36,24,48,26,17], instance segmentation [1,45], and panoptic segmentation [27,41] problems. According to the kinds of supervision, these approaches use image-level [16,36,24,48,26,41], point-level [2,38], scribble-level [47,31], or box-level [43,45,9] labels for training. Among them, image-level label-based approaches are the most prevalent. These approaches generally rely on the CAM [55,40] to extract spatial information from classification models, which are trained by the image-level labels. Though great progress has been achieved by these approaches on the semantic segmentation task, it is generally hard to distinguish different instances of the same class with only image-level labels, especially on large-scale datasets with many overlapping instances. Li et al. [27] proposes to address this problem by additionally using bounding-box annotations, which however takes much more time to annotate. PanopticFCN [29] alternatively proposes to use coarse polygons to supervise the panoptic segmentation model, which are obtained by connecting multiple point annotations for each target. PSIS [7] proposes to address the instance segmentation problem by using sparsely sampled foreground and background points in each bounding box. Though these approaches achieve better results, their annotation burden is significantly heavier than image-level labels. In this paper, we try to use a new form of weak annotation for panoptic segmentation, i.e., a single point for each target. We demonstrate that this supervision can achieve competitive performance compared with previous approaches while significantly reducing the annotation burden.

2.3 Point-Level Labels in Visual Tasks

Recently point-level annotation has drawn interest in a broad range of computer vision tasks. Beside the works concerning the detection and segmentation tasks [7,29,38,2], some works adopt point-level labels to train crowd counting [50,33] models. SPTS [37] proposes to use points for the text spotting problem. Chen et al. [5] propose addressing weakly-supervised detection problems using point labels. Besides, point labels also play an essential role in interactive segmentation models, where users provide interactive hints through point-level clicks [53,35,42]. To the best of our knowledge, there are still no approaches to training panoptic segmentation models using only a single point per target.

3 Approach

In this section, we elaborate on the details of the proposed approach. Fig. 2 illustrates the overall framework, which can be decomposed into two major components, a label generation model and a panoptic segmentation model. These two components share the same backbone and the transformer encoder [56]. The

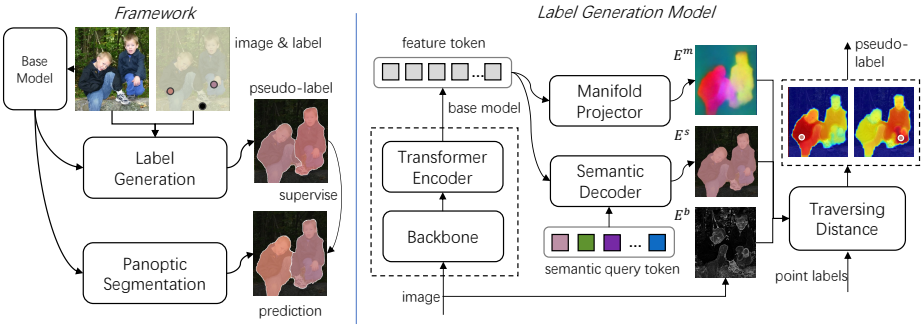

Fig. 2. Illustration of the proposed approach. Left: the overall framework, which contains a label generation model and a panoptic segmentation model. The former produces dense panoptic pseudo-labels from point-level labels, the latter is responsible for the final panoptic segmentation prediction. Right: the detailed pipeline of the label generation model. Please see Sec. 3 for more information. Best viewed in color.

label generation model is the core of the weakly-supervised learning, which is responsible for obtaining dense panoptic pseudo-masks from weak point-level labels. The panoptic segmentation model is the same as those fully-supervised ones and learns from the panoptic pseudo-masks. All these models are trained as a whole in an end-to-end manner. After the training stage, the label generation model can be removed, only leaving a standard panoptic segmentation model for usage. Hence, it does not bring any computation or memory overhead than other fully-supervised methods.

3.1 Dense Semantics from Point Supervision

Semantic parsing is the cornerstone of our approach to producing dense pixel-level pseudo-masks from sparse point-level labels. We decompose the panoptic pseudo-label generation problem into two steps: semantic parsing and instance discrimination. In the semantic parsing step, the semantic probabilities for all the pixels in the image are first obtained. By means of this, the latter problem could be reduced to partition pixels within each class into different instances, reducing the solution space and improving the estimation robustness.

To generate semantic segmentation results, we adopt a set of semantic query tokens, which has a one-to-one correspondence to the semantic classes, as shown in Fig. 2. The semantic decoder is made of transformer decoder layers following the Panoptic SegFormer [30]. It contains a mask branch to decode masks from tokens and a classification branch to decode the class probabilities. The semantic segmentation probabilities are then obtained by multiplying together the mask probabilities and the class probabilities.

Let $P \in \mathbb{R}^{HW \times C}$ denote the semantic segmentation probabilities of the C classes. Given a set of N point-level labels, it can be mapped to a set of N^s labeled semantic pixels, $Y = \{(x_i, c_i)\}_{i=1}^{N^s}$, where x_i and c_i are the coordinate

and class index of the ith pixel, respectively. The mapping could be implemented by coloring the surrounding pixels of each point-level label and applying the same geometric augmentations as the input image. The partial cross-entropy loss for semantic segmentation is the average on labeled pixels:

$$\mathcal{L}_{sem} = -\frac{1}{N^s} \sum_{(x_i,c_i)\in Y} \log P_{x_i,c_i}, \tag{1}$$

To supplement the sparse partial cross-entropy loss, inspired by [45], we adopt the image texture-based constraints densely on all the pixels, a.k.a., color-prior loss. Let $P_i, P_j \in \mathbb{R}^C$ denote the class probabilities of the ith and jth pixels, respectively. The color prior loss is defined as:

$$\mathcal{L}_{col} = -\frac{1}{Z} \sum_{i=1}^{HW} \sum_{j\in\mathcal{N}(i)} A_{i,j} \log P_i^T P_j, \tag{2}$$

where $P_i^T P_j$ measures the predicted probability similarity of the pair, higher values indicate that the prediction of the pair tends to be the same class. $A_{i,j}$ is the color-prior affinity following [45], which is obtained by thresholding the pixel similarity computed in the LAB color space with threshold 0.3. $\mathcal{N}(i)$ is the set of neighbor pixel indices of i. $Z = \sum_{i=1}^{HW} \sum_{j\in\mathcal{N}(i)} A_{i,j}$ is the normalization factor. By optimizing Eq. 2, neighboring pixels owning similar colors are encouraged to derive the same semantic prediction. Experiments in Sec. 4.3 demonstrate this strategy can effectively improve the mask quality.

3.2 Traversing Distance and Mask Generation

After obtaining the semantic classes of each pixel, the challenge of generating panoptic masks is mainly reduced to discriminating different instances in the same class. We propose a principled approach to address this problem by assigning each pixel to the nearest point label, where the distance is defined by the proposed traversing distance, as illustrated in Fig. 3.

Denote the cost of traveling from pixel i to point label s by $\mathcal{D}_{i,s}$, the target is to find the nearest point label $\hat{s}$, and mark pixel i as the foreground of the corresponding segmentation target. Formally,

$$\mathcal{D}_{i,s} = \min_{\Gamma_{i,s}} \int E(x)\Gamma_{i,s}(x)dx, \tag{3}$$

where, $\Gamma_{i,s}$ is a path from pixel i to point label s. E describes the non-negative traversing cost along the path. In discrete digital images, the path is defined as a sequence of continual pixels under the 8-neighborhood connection.

In this framework, the question reduces to defining proper transferring cost E to help distinguish different instances. We consider three criteria to accomplish this task: semantic similarity, low-level boundary cues, and high-level instance-aware manifold knowledge. For clarity, we slightly abuse the notation to denote

Fig. 3. (a) Illustration of the traversing distance method. Each pixel finds the point label with the minimum traversing cost and is assigned to the corresponding target. (b) Example of a two-instance case, showing the traversing cost of each pixel to each point label. Highlighted regions have low costs. Please see Sec. 3.2 for details. Best viewed in color.

the cost between neighboring pixels i and j by $E_{i,j}$, which is composed of the three items:

$$E_{i,j} = E^s_{i,j} + \lambda_b E^b_{i,j} + \lambda_m E^m_{i,j}, \tag{4}$$

where, λ_b and λ_m are the weights controlling the relative importance.

$E^s_{i,j}$ reflects the semantic similarity, whose value is low if the neighboring pixels belong to the same class. It is defined by the aforementioned semantic probabilities:

$$E^s_{i,j} = \sum_{c=1}^{C} |P_{i,c} - P_{j,c}|, \tag{5}$$

where, $P_{i,c}$ is the probability of the ith pixel belonging to class c. By means of this, crossing pixels of different classes are costly. Thus, paths residing in the object interiors are encouraged. This strategy could help pixels be assigned to the point within a coherent class region rather than geometrically closer points but with different classes.

$E^b_{i,j}$ defines the boundary cost considering the low-level image textures. Given the boundary map $B \in \mathbb{R}^{HW}$ obtained by edge filters, $E^b_{i,j}$ is counted by the non-negative value at the target location. In this way, the paths are encouraged not to cross the boundaries:

$$E^b_{i,j} = |B_j|, \tag{6}$$

In this paper, we adopt the efficient Sobel filter to compute the boundary map. The boundary cost implicitly assumes that regions of coherent colors are more likely to belong to the same class and instance, which has been proved experimentally by previous works [24,25,45] in addressing the segmentation tasks.

$E^m_{i,j}$ provides high-level knowledge to distinguish instances, which is learned by the deep model online. We add a manifold projector to the transformer encoder to produce dense features to compute the instance similarity, as illustrated

in Fig. 2. The manifold projector firstly reshapes the feature tokens back to 2D spatial features. Features from different pyramid levels are bilinearly sampled to the same size and summed together. Then, the projection is obtained by a 2-layer MLP model, implemented by two 1×1 convolution layers interleaved by a ReLU activation. Given the L2-normalized feature map $F \in \mathbb{R}^{HW \times D}$ produced by the manifold projector, the cost is defined by the non-negative cosine distance:

$$E^m_{i,j} = \max\{1 - F_i^T F_j, 0\}, \tag{7}$$

where, the projected features belonging to the same instance are similar and produce low costs. In this way, paths are encouraged not to cross different instances. To convey instance-aware knowledge, the manifold projection model should be trained by instance-aware constraints. For clarity, we postpone explaining the details in the Sec. 3.3.

After obtaining the criteria, the next step is finding the shortest path between each pair of pixels and point labels in the 8-neighborhood graph. Note that the graph is very sparse because each pixel only connects to its local neighbors. Let N denote the total number of point labels and M denote the number of pixels. Then, the minimum distance $\mathcal{D}_{i,s}$ in Eq. 3 can be efficiently solved by the shortest distance algorithm with time complexity $O(MN \log M)$.

It is noteworthy that the distance measurement in Eq. 3 can only produce connected components. As a result, if an instance is overlapped and separated into several parts by some region belonging to different classes, the farther parts would be assigned with the wrong class. To overcome this limitation, we use the class compatibility between the pixel and the point label to reweight the distance before assigning pixels to point labels.

$$\hat{s} = \text{argmin}_s \left[\left(\hat{\mathcal{D}}_{i,s} - 1 \right) \cdot \left(P_i^T P_s \right) \right], \tag{8}$$

where, P_i is the probability of the ith pixel obtained by the semantic segmentation branch. P_s is the one-hot encoding of the point label's ground truth class. $\hat{\mathcal{D}}$ is the normalized version of $\mathcal{D}$ in range $[0, 1]$. Finally, we judge that pixel i is part of the instance holding point label $\hat{s}$ according to Eq. 8, and the whole image is parsed into non-overlapping regions, obtaining panoptic pseudo-masks.

3.3 Weakly-Supervised Training

In this section, we elaborate on training the whole model with the generated pseudo-masks. Firstly, the panoptic segmentation model, as illustrated in Fig. 2, is trained by the pseudo-masks. As aforementioned, we adopt the Panoptic SegFormer architecture, which contains a localization decoder to help quickly converge to the target locations, a classification branch to predict class probabilities for each query token, and a mask decoder to decode masks. They are optimized by the localization loss, the focal loss, and the dice loss, respectively. For simplicity, here we denote all these losses to train the panoptic segmentation model as $\mathcal{L}_{pan}$. Please refer to the paper [30] for more details.

In addition to the panoptic segmentation model, the manifold projector used in Eq. 7 also needs to train to provide instance-aware representations. Here we utilize a contrastive learning strategy [19,6] to optimize the manifolds with the pseudo-masks. Let $\mathcal{M}_i \in \mathbb{R}^{HW}$ denote the pseudo-mask of target i that contains point label with coordinate x_i. The global representation of target i is the masked average of the projection $F \in \mathbb{R}^{HW \times D}$:

$$\bar{F}_i = \frac{1}{Z_i} \sum_{j=1}^{HW} \mathcal{M}_{i,j} F_j, \tag{9}$$

where, $Z_i = \sum_{j=1}^{HW} \mathcal{M}_{i,j}$ is for normalization, $F_j \in \mathbb{R}^D$ is the feature at pixel j.

The loss for the projector is the average of all point-to-target contrasts:

$$\mathcal{L}_{cl} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(F_{x_i}^T \bar{F}_i / \tau\right)}{\sum_{j=1}^{N} \exp\left(F_{x_i}^T \bar{F}_j / \tau\right)}, \tag{10}$$

where, N is the total number of point labels, τ is a scale factor and is set 0.07 following [19]. When the setting is extended to annotate each target with multiple points, the sum in the denominator of the contrasts should iterate over the number of targets. With this optimization target, the feature projections at the labeled points are encouraged to be coherent within the estimated instance region and distinctive from the others. The pseudo-mask estimation and the manifold projector mutually benefit each other and improve the model together.

Taking all the above components together, our final model is end-to-end:

$$\mathcal{L}_{all} = \mathcal{L}_{sem} + \mathcal{L}_{col} + \mathcal{L}_{pan} + \mathcal{L}_{cl}, \tag{11}$$

After training, only the panoptic segmentation model is kept for testing, thus not incurring any computation or memory overhead compared with previous fully-supervised panoptic segmentation models.

4 Experiments

In this section, we discuss the experiments. We first describe the experiment setting and implementation details and then elaborate on the analyses and comparison using the VOC and COCO datasets.

4.1 Datasets

Pascal VOC dataset [13,18] contains 10582 training images and 1449 validation images. It has 20 thing classes and one stuff class. By default, a single point label per target is sampled with the uniform distribution from the masks, which is fixed through all the experiments. To study the influence of point label distribution, we also adopt other sampling strategies in Sec. 4.3 for analyses. We analyze our approach with the panoptic quality (PQ), segmentation quality (SQ), recognition quality (RQ), and intersection over union (IoU) metrics.

MS COCO dataset [32] contains 118k images for training, 5k images for validation, and 20k images for testing. It contains 80 thing classes and 53 stuff classes. The same point sampling strategy and metrics are applied to the COCO.

4.2 Implementation Details

Architecture. We base our approach on the Panoptic SegFormer with a ResNet50 backbone [20]. As mentioned in Sec. 4.3, we adopt an extra group of semantic query tokens to produce the semantic segmentation results. Specifically, the C semantic tokens produce C masks and classification scores, which are multiplied together and projected by a linear layer followed by a Softmax function to produce semantic probabilities. The mask decoder contains 6 transformer decoder layers [30], which has the same architecture as the panoptic segmentation model. The color prior loss in Eq. 2 is constructed by sampling neighboring pixels with kernel size 5 and dilation rate 2, and the loss is amplified by factor 3. The image edge used in Eq. 6 is obtained by the Sobel filter in the LAB color space and its absolute values are normalized into the range $[0, 1]$. The panoptic segmentation model follows the same setting as [30], and the number of query tokens for panoptic segmentation is set to 300.

Optimization. We follow previous practice [30] for training, i.e., AdamW optimizer with learning rate 1.4×10^{-4}, weight decay 1.4×10^{-4}, and batch size 8. The learning rates for the backbone parameters are multiplied by the factor 0.1. To stabilize the training, we adopt a linear warm-up strategy for the losses $\mathcal{L}_{pan}$ and $\mathcal{L}_{cl}$ during the first training epoch, so that reliable pseudo-panoptic masks can be obtained, which depends on the well-learned semantic parsing results. The balancing weights λ_b and λ_m in Eq. 4 are all set 0.1 by default. In experiments, we extend each point label to a square region with a size of 17 pixels to facilitate the training of semantic segmentation, as explained in Eq. 1. The shorter sizes of input images are resized to 600 and 800 on the VOC and COCO, respectively. On the VOC, we train 20 epochs and decay the learning rate with a factor of 0.1 after epoch 15. On the COCO, we follow the $1\times$ schedule to train 12 epochs and decay the learning rate after epoch 8.

4.3 Ablation Studies

In this section, we conduct analytic experiments on the VOC dataset to reveal the properties of the proposed method with point-level supervision.

Instance Discrimination. We first conduct experiments to demonstrate the effectiveness of the proposed traversing distance-based instance discrimination approach. Tab. 1 shows the ablation results of the distance measurement criteria in Eq. 4. With only the semantic probabilities, the model achieves PQ 46.6% on the VOC val set. The boundary criterion and the manifold criterion improve the

Table 1. Ablation studies for the proposed traversing distance-based instance discrimination. Results are reported on the VOC val set. E^s, E^b, and E^m denote the criteria of semantic similarity, low-level boundary, and high-level manifold, respectively.

E^s	E^b	E^m	PQ	PQ^{th}	PQ^{st}
✓			46.6	44.5	89.1
✓	✓		48.5	46.5	89.3
✓		✓	49.4	47.4	89.2
✓	✓	✓	**49.8**	**47.8**	**89.5**

Table 2. Influence of the hyper-parameters in computing the traversing cost. Results are reported on the VOC val set. We conduct experiments by fixing one hyper-parameter and alter another.

λ_b	λ_m	PQ	PQ^{th}	PQ^{st}
0.0	0.1	49.4	47.4	89.2
0.1	0.1	**49.8**	**47.8**	**89.5**
0.5	0.1	48.3	46.3	89.1
1.0	0.1	48.4	46.3	89.4
0.1	0.0	48.5	46.5	89.3
0.1	0.1	**49.8**	**47.8**	**89.5**
0.1	0.5	49.3	47.3	89.2
0.1	1.0	48.9	47.0	89.1

baseline result to 48.5% and 49.4%, respectively. We noticed that this improvement is mainly due to the PQ^{th}, which is improved from 44.5% to 46.5% and 47.4%, respectively. And the results of stuff classes are relatively similar, demonstrating that the low-level image cues and high-level instance-aware manifold can effectively help to identify different instances. Finally, with all the criteria, our approach achieves the final result of PQ 49.8% and PQ^{th} 47.8%.

Hyper-parameter Sensitivity. We conduct experiments to study the sensitivity of the hyper-parameters used in Eq. 4. To save the search cost, we fix one parameter and adjust another. Results reported in Tab. 2 show that the scale of the additional criteria for the instance discrimination, i.e., the boundary and the feature manifold criteria, should be approximately one order of magnitude smaller than the semantic criterion. It is noteworthy that even larger values are not optimal, they still boost the baseline's performance from PQ 46.6% to 48.3% or higher, demonstrating the robustness of the proposed approach.

Point Sampling Strategies. We conduct experiments to study the influence of the position distribution bias of the point labels. In addition to the uniform sampling strategy, we also tried the center-biased and the border-biased sampling strategies. Specifically, we first compute the euclidean distance of each pixel to the centroid of the corresponding ground truth mask. Then, we build the probability density map according to the square of the euclidean distance. Finally,

Table 3. Influence of the point sampling strategy. Results are reported on the VOC val set. "Border" and "Center" refer to strategies that prefer target border regions and center regions, respectively.

Method	mIoU	PQ	SQ	RQ
Border	65.6	44.7	78.1	55.7
Uniform	67.5	49.8	78.4	62.0
Center	**67.7**	**50.9**	**79.1**	**62.8**

Table 4. Ablation study of the semantic segmentation submodule. Results are reported on the VOC val set.

Method	mIoU	PQ	SQ	RQ
w/o $\mathcal{L}_{col}$	62.2	41.3	74.3	54.0
w/ $\mathcal{L}_{col}$	67.5	49.8	78.4	62.0

the points are sampled based on the normalized probability for each target to obtain border-biased labels. The center-biased labels are sampled in a similar way by reversing the probabilities. Results in Tab. 3 show that the center-biased labels achieve the best performance, and the border-biased labels perform worst. While the SQ values of the three methods are relatively similar, the RQ of the border-biased strategy is much worse than the others, revealing that annotations near borders are harmful to discriminating different targets, while annotations at center regions provide more robust results. This phenomenon has positive meanings because center annotation accords with human intuition and is easier in practice.

Semantic Segmentation Module. In this section, we conduct ablation experiments to study the semantic segmentation submodule. Results in Tab. 4 demonstrate the low-level cues can effectively improve the segmentation performance from mIoU 62.2% to mIoU 67.5%. The improvement of the semantic segmentation quality not only improves the quality of the panoptic masks, i.e., SQ is improved from 74.3% to 78.4%, but also benefits the localization of the targets, i.e., a significant improvement of the RQ from 54.0% to 62.0%. We conjecture the reason is that more accurate semantic probabilities provide more reliable instance discrimination cues for Eq. 4 to distinguish different targets.

4.4 Comparison with Related Works

We compare our approach with the related works in Tab. 5. Compared to the JTSM [41] that uses image-level labels, our approach achieves significant improvement with the help of point-level labels, which improves the PQ with +10.8% on the VOC, and +24.0% on the COCO. It demonstrates that point-level labels are promising in addressing panoptic segmentation problems. As pointed out by Bearman et al. [2], the point-level labels only cost marginally above image-level labels. For example, on the VOC dataset, image labels cost

Table 5. Comparison with related works on the VOC and the COCO datasets. Results are reported on the COCO val set and the VOC val set. $\mathcal{M}$ mask annotation for fully-supervised learning, $\mathcal{B}$ bounding-box level supervision, $\mathcal{I}$ image-level supervision, $\mathcal{P}$ the proposed point-level supervision, $\mathcal{P}_{10}$ point-level supervision with 10 points per target.

Method	Backbone	Supervision	COCO			VOC		
			PQ	PQ^{th}	PQ^{st}	PQ	PQ^{th}	PQ^{st}
Panoptic FPN [22]	R50	$\mathcal{M}$	41.5	48.3	31.2	65.7	64.5	90.8
K-Net [54]	R50	$\mathcal{M}$	47.1	51.7	40.3	-	-	-
MaskFormer [8]	R50	$\mathcal{M}$	46.5	51.0	39.8	-	-	-
Panoptic SegFormer [30]	R50	$\mathcal{M}$	48.0	52.3	41.5	67.9	66.6	92.7
Li et.al. [27]	R101	$\mathcal{B}+\mathcal{I}$	-	-	-	59.0	-	-
JTSM [41]	R18-WS	$\mathcal{I}$	5.3	8.4	0.7	39.0	37.1	77.7
PanopticFCN [29]	R50	$\mathcal{P}_{10}$	31.2	35.7	24.3	48.0	46.2	85.2
Ours	R50	$\mathcal{P}$	29.3	29.3	29.4	49.8	47.8	89.5
Ours	R50	$\mathcal{P}_{10}$	33.1	33.6	32.2	56.6	54.8	91.4

on average 20.0 sec/img, while point labels cost 22.1 sec/img, where the difference is marginal compared with the full labels' 239.7 sec/img. Compared to the PanopticFCN [29] using ten points to connect polygons for training, our approach achieves competitive performance when using only a single point as annotation, which is +1.8% on the VOC and -1.9% on the COCO in respect to the PQ. Note that these comparable results are achieved by using only 1/10 of the annotations of the PanopticFCN. When increasing the annotation to ten points per target, our approach achieves +8.6% and +1.9% improvements on the VOC and the COCO datasets compared with the PanopticFCN, demonstrating the scalability of our approach in utilizing more points per target.

To help qualitatively study the results, we visualize the predictions on the val set of VOC and COCO. Results in Fig. 4 show that our approach performs generally well in handling scenes with complex multiple instances and classes. We also show the results for hard examples, which contain extremely many instances with small scales. In these cases, some instances are missing in the prediction. Improving the performance with these small and thin objects when only accessing point-level labels may be a potential direction in future studies.

5 Conclusions

In this paper, we propose a new paradigm for weakly-supervised panoptic segmentation using a single point label for each target. To tackle this problem, we propose a principled approach that generates panoptic pseudo-masks by solving the minimization problem of pixel-to-point traversing costs, which integrates semantic similarity, low-level texture cues, and high-level knowledge to distinguish different targets. We demonstrate its effectiveness and study the influence of point labels through analytical experiments. Besides, we achieve new state-of-the-art performance with point-level labels on the VOC and COCO datasets.

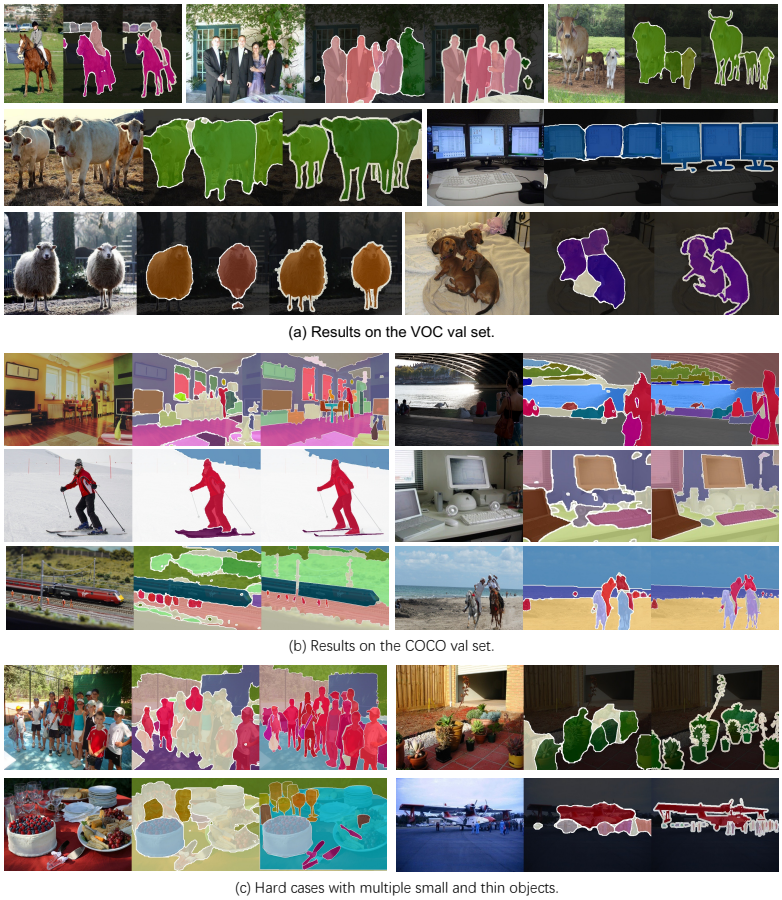

Fig. 4. Visualization of the panoptic segmentation results. The models are trained with a *single point* per target as annotation. Each group from left to right are the input, prediction, and ground truth, respectively. Best viewed in color.

Acknowledgments This work was supported in part by the Major Project for New Generation of AI (No.2018AAA0100400), the National Natural Science Foundation of China (No. 61836014, No. U21B2042, No. 62072457, No. 62006231).

References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2209–2218 (2019)
2. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. In: Proceedings of the European conference on computer vision. pp. 549–565. Springer (2016)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proceedings of the European conference on computer vision. pp. 213–229. Springer (2020)
4. Chan, L., Hosseini, M.S., Plataniotis, K.N.: A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. International Journal of Computer Vision **129**(2), 361–384 (2021)
5. Chen, L., Yang, T., Zhang, X., Zhang, W., Sun, J.: Points as queries: Weakly semi-supervised object detection by points. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8823–8832 (2021)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
7. Cheng, B., Parkhi, O., Kirillov, A.: Pointly-supervised instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2617–2626 (2022)
8. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems **34** (2021)
9. 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)
10. De Geus, D., Meletis, P., Dubbelman, G.: Panoptic segmentation with a joint semantic and instance segmentation network. arXiv preprint arXiv:1809.02110 (2018)
11. 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)
12. 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. ICLR (2021)
13. 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**(2), 303–338 (2010)
14. Fan, J., Zhang, Z., Song, C., Tan, T.: Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4283–4292 (2020)
15. Fan, J., Zhang, Z., Tan, T.: Employing multi-estimations for weakly-supervised semantic segmentation. In: Proceedings of the European Conference on Computer Vision (2020)
16. 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)

17. Fan, R., Hou, Q., Cheng, M.M., Yu, G., Martin, R.R., Hu, S.M.: Associating inter-image salient instances for weakly supervised semantic segmentation. In: Proceedings of the European Conference on Computer Vision. pp. 367–383 (2018)
18. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: Proceedings of the International Conference on Computer Vision. pp. 991–998. IEEE (2011)
19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
20. 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)
21. 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)
22. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6399–6408 (2019)
23. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9404–9413 (2019)
24. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: Proceedings of the European Conference on Computer Vision. pp. 695–711. Springer (2016)
25. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in Neural Information Processing Systems. pp. 109–117 (2011)
26. Lee, J., Kim, E., Lee, S., Lee, J., Yoon, S.: Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5267–5276 (2019)
27. Li, Q., Arnab, A., Torr, P.H.: Weakly-and semi-supervised panoptic segmentation. In: Proceedings of the European Conference on Computer Vision. pp. 102–118 (2018)
28. Li, Q., Qi, X., Torr, P.H.: Unifying training and inference for panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13320–13328 (2020)
29. Li, Y., Zhao, H., Qi, X., Chen, Y., Qi, L., Wang, L., Li, Z., Sun, J., Jia, J.: Fully convolutional networks for panoptic segmentation with point-based supervision. arXiv preprint arXiv:2108.07682 (2021)
30. Li, Z., Wang, W., Xie, E., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., Lu, T.: Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1280–1289 (2022)
31. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3159–3167 (2016)
32. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755. Springer (2014)

33. Liu, Y., Xu, D., Ren, S., Wu, H., Cai, H., He, S.: Fine-grained domain adaptive crowd counting via point-derived segmentation. arXiv preprint arXiv:2108.02980 (2021)
34. 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)
35. Maninis, K.K., Caelles, S., Pont-Tuset, J., Van Gool, L.: Deep extreme cut: From extreme points to object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 616–625 (2018)
36. Pathak, D., Krähenbühl, 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)
37. Peng, D., Wang, X., Liu, Y., Zhang, J., Huang, M., Lai, S., Zhu, S., Li, J., Lin, D., Shen, C., et al.: Spts: Single-point text spotting. arXiv preprint arXiv:2112.07917 (2021)
38. Qian, R., Wei, Y., Shi, H., Li, J., Liu, J., Huang, T.: Weakly supervised scene parsing with point-based distance metric learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8843–8850 (2019)
39. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
40. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
41. Shen, Y., Cao, L., Chen, Z., Lian, F., Zhang, B., Su, C., Wu, Y., Huang, F., Ji, R.: Toward joint thing-and-stuff mining for weakly supervised panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16694–16705 (2021)
42. Sofiiuk, K., Petrov, I.A., Konushin, A.: Reviving iterative training with mask guidance for interactive segmentation. arXiv preprint arXiv:2102.06583 (2021)
43. 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 Conference on Computer Vision and Pattern Recognition. pp. 3136–3145 (2019)
44. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7262–7272 (2021)
45. Tian, Z., Shen, C., Wang, X., Chen, H.: Boxinst: High-performance instance segmentation with box annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5443–5452 (2021)
46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
47. Vernaza, P., Chandraker, M.: Learning random-walk label propagation for weakly-supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 3, p. 3 (2017)
48. Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S.: Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7268–7277 (2018)

49. Wu, Y., Zhang, G., Xu, H., Liang, X., Lin, L.: Auto-panoptic: Cooperative multi-component architecture search for panoptic segmentation. Advances in Neural Information Processing Systems **33**, 20508–20519 (2020)
50. Zand, M., Damirchi, H., Farley, A., Molahasani, M., Greenspan, M., Etemad, A.: Multiscale crowd counting and localization by multitask point supervision. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1820–1824. IEEE (2022)
51. Zhang, D., Han, J., Cheng, G., Yang, M.H.: Weakly supervised object localization and detection: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
52. Zhang, M., Zhou, Y., Zhao, J., Man, Y., Liu, B., Yao, R.: A survey of semi-and weakly supervised semantic segmentation of images. Artificial Intelligence Review **53**(6), 4259–4288 (2020)
53. Zhang, S., Liew, J.H., Wei, Y., Wei, S., Zhao, Y.: Interactive object segmentation with inside-outside guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12234–12244 (2020)
54. Zhang, W., Pang, J., Chen, K., Loy, C.C.: K-net: Towards unified image segmentation. Advances in Neural Information Processing Systems **34** (2021)
55. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929 (2016)
56. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: Proceedings of the International Conference on Learning Representations (2021)