Exclusivity-Guided Mask Learning for Semi-Supervised Crowd Instance Segmentation and Counting

Jiyang Huang[1], Hongru Chen[1], Wei Lin[2], Jia Wan[1]*, and Antoni B. Chan[2]

[1] Harbin Institute of Technology, Shenzhen, China
[2] City University of Hong Kong, Kowloon Tong, Hong Kong
{jiyanghuang0127,elonlin24,jiawan1998}@gmail.com
{24S151003}@hit.edu.cn
{abchan}@cityu.edu.hk

Abstract. Semi-supervised crowd analysis is a prominent area of research, as unlabeled data are typically abundant and inexpensive to obtain. However, traditional point-based annotations constrain performance because individual regions are inherently ambiguous, and consequently, learning fine-grained structural semantics from sparse annotations remains an unresolved challenge. In this paper, we first propose an Exclusion-Constrained Dual-Prompt SAM (EDP-SAM), based on our Nearest Neighbor Exclusion Circle (NNEC) constraint, to generate mask supervision for current datasets. With the aim of segmenting individuals in dense scenes, we then propose Exclusivity-Guided Mask Learning (XMask), which enforces spatial separation through a discriminative mask objective. Gaussian smoothing and a differentiable center sampling strategy are utilized to improve feature continuity and training stability. Building on XMask, we present a semi-supervised crowd counting framework that uses instance mask priors as pseudo-labels, which contain richer shape information than traditional point cues. Extensive experiments on the ShanghaiTech A, UCF-QNRF, and JHU++ datasets (using 5%, 10%, and 40% labeled data) verify that our end-to-end model achieves state-of-the-art semi-supervised segmentation and counting performance, effectively bridging the gap between counting and instance segmentation within a unified framework. Code can be found at https://github.com/JoyceeH0127/ECCV2026XMask.

Keywords: Semi-supervised learning · Crowd Counting · Instance Segmentation

1 Introduction

Crowd scene analysis has long been a pivotal research topic in computer vision, driven by its critical applications in public safety, smart city management, and

* Corresponding author

traffic monitoring. Despite significant progress in density estimation and localization [18, 25, 39, 49, 65], the field remains constrained by a severe data annotation bottleneck. The prohibitive cost of acquiring pixel-level ground truth forces a reliance on point-based annotations, which inherently lack structural semantics regarding individual shape and occupancy.

Current crowd analysis research spans multiple directions: density regression and point-based localization, which treat individuals as dimensionless points, ignoring spatial extent; and object detection research [11, 14, 33, 51, 56] struggles with bounding box regression in scenarios of severe occlusion. Other related methods [37, 45] connect the detection task with counting, but lack in semantic information. While recent foundation models like SAM [21] demonstrate zero-shot segmentation potential, their direct application is hindered by high computational latency and performance degradation in dense, ambiguous regions [7]. Consequently, learning fine-grained structural semantics from sparse annotations remains an open problem. In particular, how to harness additional data to compensate for the lack of dense supervision has not been sufficiently investigated.

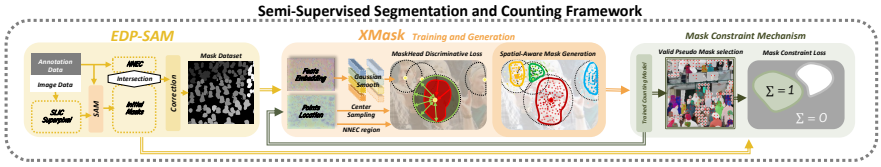

Fig. 1: Conception of our approach. EDP-SAM is designed to generate data with mask supervision. To acquire segmentation capabilities, we propose a semi-supervised XMask method to learn a discriminative feature representation. Subsequently, we present a mask-constraint strategy composing the final step in our framework, coupling semantic and spatial information to enhance counting accuracy.

To address these limitations, we propose a novel paradigm that extends crowd counting to dense instance segmentation without relying on expensive point and mask annotations, as illustrated in Fig. 1. Our objective is to achieve high instance segmentation performance while enhancing counting accuracy in dense crowds. First, we employ SAM [21] to generate datasets with masks by using existing annotations as point prompts and SLIC [2]-generated superpixels for region guidance. The **N**earest-**N**eighbor **E**xclusion **C**ircle **(NNEC)** is further introduced as a spatial constraint to suppress interference from adjacent instances. This design, **E**xclusion-Constrained **D**ual-**P**rompt **S**AM **(EDP-SAM)**, yields spatially consistent and reliable pseudo-masks. After manual correction, the resulting datasets are subsequently used for model training and standardized comparison with baseline methods.

We then propose an E**X**clusivity-Guided **Mask** Learning **(XMask)** for crowd instance segmentation by enforcing spatial separation through a discriminative mask objective. First, we propose NNEC, a metric-induced local topological partition that guarantees first-order exclusivity under sparse supervision. Next, by leveraging discriminative feature embedding learning [5, 9, 46], we reformu-

late the instance segmentation task as a semantic segmentation problem within dynamically defined local regions, which significantly enhances computational efficiency. Additionally, to improve feature continuity and training stability, we incorporate Depthwise Gaussian smoothing into the embedding space and use a Differentiable Center Sampling strategy to ensure gradient consistency.

Based on the proposed crowd instance segmentation architecture, we further design a semi-supervised mask-constraint mechanism that provides richer guidance than point labels for semi-supervised crowd counting. By incorporating a set of mask constraint losses and coupling them with a valid mask filtering strategy, the proposed mechanism forms a feedback loop that enhances the reliability and consistency of pseudo-mask supervision. Rather than leveraging feature embeddings to generate instance pseudo-masks for enforcing a strict spatial containment, Abousamra, Shahira et al. [1] impose a topological constraint on the likelihood map's landscape using persistence loss, ensuring the number of connected components in the predicted heatmap corresponds exactly to the number of ground truth dots. Our design allows the segmentation branch to regularize crowd counting in a semi-supervised manner. Crucially, our specialized framework delivers segmentation performance on par with foundation models (e.g., SAM) but with drastically reduced computational costs, presenting a robust and efficient alternative for dense crowd analysis. Extensive experiments on the ShanghaiTech A (ShTech A), UCF-QNRF, and JHU++ datasets (using 5%, 10%, and 40% labeled protocols) demonstrate that our end-to-end model significantly outperforms point-prompt FastSAM [66]-based methods in both inference speed and segmentation accuracy.

The contributions of our paper are summarized as follows:

- We propose Exclusion-Constrained Dual-Prompt SAM (EDP-SAM), an automatic mask generation framework for dense crowds that guides SAM with head-point and superpixel dual prompts while enforcing a Nearest Neighbor Exclusion constraint to prevent cross-instance leakage. This design improves individual separation in highly congested scenes, and the resulting pseudo-masks are lightly refined to build a high-quality crowd segmentation dataset.
- We introduce Exclusivity-Guided Mask Learning (XMask), an exclusivity-guided semi-supervised method for dense crowd instance segmentation that enforces spatial separation through a discriminative mask objective, coupled with Gaussian smoothing and differentiable center sampling for robust learning under limited supervision.
- Building on XMask, we further introduce an instance mask prior framework for semi-supervised crowd counting, in which reliable instance masks replace point annotations as structured pseudo-labels. By leveraging richer spatial and shape information instead of sparse point cues, the framework produces more accurate pseudo-labels and significantly enhances the performance.
- Our semi-supervised strategy achieves state-of-the-art results, effectively bridging the gap between crowd counting and segmentation, constructing feature structural consistency, demonstrating superior segmentation efficiency and higher recall while maintaining precise counting performance.

2 Related Works

2.1 Instance Segmentation

Instance segmentation is a core foundational task in computer vision. It not only requires detecting all objects of interest in an image but also assigns precise category labels and instance IDs to each object's pixels.

Crowd Instance Segmentation. Crowd instance segmentation aims to separate each individual in highly congested scenes, where severe occlusion and overlap make precise delineation challenging. Early deep learning approaches [19] relied on fully convolutional networks to directly predict crowd regions. With the rise of foundation models such as SAM [21] and FastSAM [66], recent efforts have shifted toward adapting prompt-driven segmentation frameworks to dense small-object scenarios. For instance, CrowdSAM [7] utilizes density-derived prompts and post-processing refinement to enhance mask quality. Nevertheless, limitations in inference efficiency and scalability remain unresolved.

Two Stage Instance Segmentation. Two-stage instance segmentation follows a proposal-to-refinement pipeline: it first generates candidate regions and then refines them into final masks. In this framework, methods are mainly categorized into top-down and bottom-up approaches [41].

Top-down instance segmentation methods rely on high-performance detectors to generate RoIs, followed by independent mask prediction within each region, effectively decomposing the task into local binary segmentation. Mask R-CNN [16] established this paradigm by adding a parallel mask branch to Faster R-CNN [44]. Subsequent works improved feature propagation and mask quality, including PANet [32] with bottom-up path aggregation, Cascade Mask R-CNN [12] with multi-stage refinement, and several boundary-aware extensions [10, 20, 62] that enhance pixel-level mask accuracy.

Bottom-up approaches [26, 31] learn pixel-wise embeddings instead of relying on predefined anchors, encouraging pixels of the same instance to cluster in a high-dimensional space while separating different instances. This paradigm is typically optimized via discriminative losses with pull–push forces. Representative works [5, 9, 26, 46]jointly predict semantic maps and pixel embeddings to simplify training, while others [3, 60] exploit spatial affinity cues for grouping. Embedding-based methods have shown particular advantages in irregular scenarios, such as biomedical images, as demonstrated by EmbedSeg [22].

Semi-supervised Instance Segmentation. This task learns from scarce annotations and abundant unlabeled data, commonly relying on pseudo-labeling or consistency constraints to exploit unlabeled samples.

Early semi-supervised instance segmentation methods were primarily built upon anchor-based frameworks such as Mask R-CNN [16]. Noisy Boundaries [57] and PAIS [17] enhanced pseudo-label robustness via boundary-aware modeling and dynamic alignment, respectively, yet remained limited by architectural complexity or label reweighting strategies. To adapt to anchor-free paradigms, Polite Teacher [13] adopted a CenterMask-style design with mutual learning, improving efficiency but showing weakness on extremely small objects. More recently,

Guided Distillation [4] introduced Vision Transformers to leverage pretrained representations, albeit with increased memory overhead.

The emergence of vision foundation models like SAM [21] has shifted semi-supervised instance segmentation toward prompt-driven and more generalizable paradigms, improving pseudo-label quality. Methods like Sˆ4M [61] and SemiSAM+ [64] exploit SAM [21] as a powerful teacher to refine mask supervision. Meanwhile, diffusion-based approaches [53, 58, 59], including DiffusionInst [15], model segmentation as a denoising process and leverage generative modeling for semantic alignment. Despite enhanced robustness to ambiguous boundaries, these methods often suffer from substantial computational overhead in dense scenarios.

2.2 Semi-supervised Crowd Counting

Semi-Supervised Crowd Counting is a specialized sub-field of computer vision that addresses the challenge of estimating the number of individuals in dense crowds using a limited amount of labeled data alongside a large corpus of unlabeled data.

Density-Based Counting. Crowd counting annotation is costly and labor-intensive. Recent Works [8, 36, 39] explore semi-supervised learning to leverage unlabeled data for improved counting performance. Liu et al. [34, 35] propose a Learning-to-Rank (L2R) module that models the ordinal relationship between image patches and their sub-patches. DACount [27] and P^3Net [28] improve density discrimination by assigning density-level labels to image patches. Approaches [40, 55] incorporate pixel-wise uncertainty to suppress noise in pseudo density maps, promoting pseudo-label quality.

Localization-Based Counting. Semi-supervised localization-based counting methods generally adopt a teacher–student framework. OT-M [29] derives pseudo-points from predicted density maps using optimal transport. CU [24], built upon P2PNet [49], improves pseudo-label reliability by selecting high-confidence patches based on uncertainty estimation. SAL [63] incorporates active learning to dynamically divide data, treating difficult samples as labeled and easier ones as unlabeled. P2R [30] introduces a points to region supervision strategy that segments regions around pseudo-points to propagate confidence more effectively than traditional P2PNet [49] matching. Consistent-Point [67] improves semi-supervised point-prompt crowd counting by refining pseudo-points through position aggregation and uncertainty calibration for more consistent supervision.

3 Methods

To ensure robust semi-supervised instance segmentation performance in dense crowds, we a propose framework that consists of three stages. As shown in Fig. 2, our model establishes a synergistic bridge between semi-supervised instance segmentation and counting within a single end-to-end architecture, enabling the two tasks to mutually reinforce each other in a closed-loop manner. The proposed

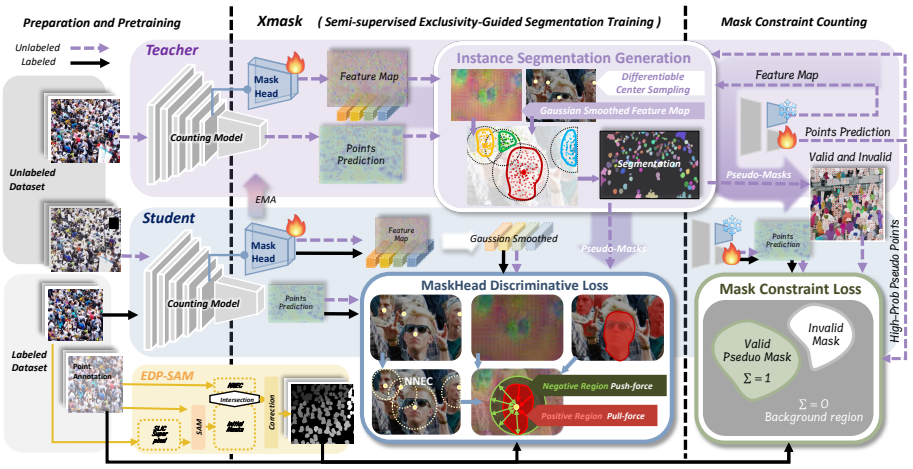

Fig. 2: Overview of the proposed framework. Our semi-supervised pipeline consists of three stages: (1) pretraining and mask preparation, (2) crowd segmentation training, and (3) mask-constrained crowd counting optimization. In Stage 1, mask annotations are generated from point labels via the proposed EDP-SAM (yellow region). Stage 2 presents the semi-supervised XMask training process, including discriminative supervision and pseudo-mask generation. The trained segmentation model is then utilized in Stage 3 to produce pseudo-masks for unlabeled data, where newly designed loss functions are introduced to further enhance counting performance via mask constraint.

method achieves state-of-the-art semi-supervised crowd instance segmentation and counting performance.

Our approach builds on P2R [30], which mitigates noisy pseudo-label locations by matching predictions within a neighborhood. Structurally, we streamline the P2PNet [49] architecture by removing the regression branch in favor of pixel grid coordinates. Our model incorporates a feature backbone, fusion neck, and classification head, alongside an innovative introduced parallel mask head for segmentation. For semi-supervised learning, we adopt the Mean Teacher paradigm [52], where the student optimizes joint losses while the teacher generates pseudo-labels, which follows standard procedure [23, 48], with parameters updated via the student exponential moving average (EMA) [6, 43]. The whole pipeline can be trained end-to-end.

3.1 Dataset Construction with Dense Mask Supervision

We introduce EDP-SAM to automatically generate and manually verify a set of high-quality masks, providing essential structural guidance for our segmentation-counting dual-tasks optimization.

Since existing crowd counting datasets only contain point annotation data $\mathbf{p}_i = (y_i, x_i)$, our primary task is to construct the corresponding mask supervision datasets. Inspired by relevant research [38, 42, 54], we also employed the SAM model for automatic construction, supplemented by partial manual correction. As illustrated in Fig. 3, we fed both the original annotation points and

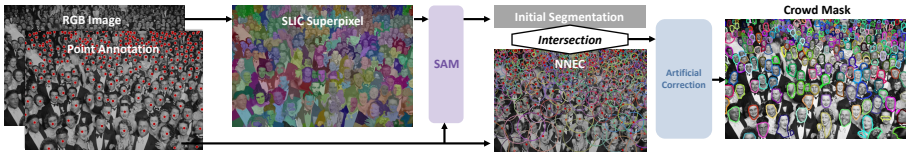

Fig. 3: Point-Superpixel Dual-Prompt SAM with NNEC Constraint Method

superpixels generated by SLIC [2] as dual-prompts into SAM. We then intersected the segmentation output from SAM with the Nearest-Neighbor Exclusion Circle (NNEC) regions corresponding to each point. The reason for our design will be illustrated in the supplementary materials.

Specifically, the purpose of applying NNEC constraint is to identify the maximum mask area corresponding to each annotation, as the SAM often struggles to segment individual shapes in dense scenes, even when specific point prompts are provided. The specific calculation of NNEC radius is as follows:

$$r_i = \min_{j \neq i} \|\mathbf{p}_i - \mathbf{p}_j\|_2, \tag{1}$$

where $\mathbf{p}_i$ and $\mathbf{p}_j$ are two point prompts indicating two head locations. Based on EDP-SAM and manual correction, we generate mask annotations for major crowd counting datasets for future crowd instance segmentation research.

3.2 Semi-supervised Crowd Instance Segmentation

Based on the mask annotations, we propose Exclusivity-Guided Mask Learning (XMask), a semi-supervised method for dense crowd instance segmentation that enforces spatial separation through a discriminative mask objective, coupled with Gaussian smoothing and differentiable center sampling for robust learning under limited supervision. As shown in Fig. 2, after obtaining a preliminary counting model, we train an instance segmentation network within the same semi-supervised framework using the proposed XMask.

MaskHead Discriminative Loss This method proposes a feature embedding learning loss function for point-supervised instance segmentation. Unlike traditional per-pixel classification paradigms, we introduce NNEC adaptive radius local circular region sampling and asymmetric marginal contrast constraints within the pixel embedding space. The whole process is visualized in Fig. 4.

Specifically, given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ and its corresponding embedding feature map $\mathbf{E} \in \mathbb{R}^{D \times H_e \times W_e}$ where D is the feature dimension, a set of model predict point annotations $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^{N}$ where $\mathbf{p}_i = (y_i, x_i)$ denotes the coordinates of annotated points, and instance-level ground truth masks $\mathbf{M} \in \mathbb{Z}_{\geq 0}^{H \times W}$.

Depthwise Gaussian Smoothing. To mitigate high-frequency interpolation artifacts and local noise in sampled feature maps, we employ a separable

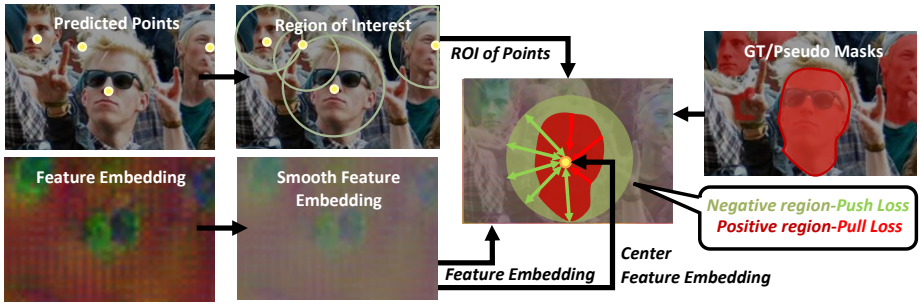

Fig. 4: MaskHead Discriminative Loss. First, use points to obtain the corresponding ROIs for computation. Based on the pseudo/ground truth mask, divide the ROIs into positive and negative regions. After applying Gaussian smoothing, extract the center feature and compute pixel-level L2 distance within feature map. Finally, applying pull and push force losses separately.

depthwise Gaussian blur that enhances spatial coherence without compromising embedding discriminability and the effectiveness is demonstrated in Sec. 4.3. Given a kernel size k and standard deviation σ, we construct discrete coordinates u and a normalized one-dimensional Gaussian kernel:

$$g(u) = \frac{\exp\left(-\frac{u^2}{2\sigma^2}\right)}{\sum_{u'} \exp\left(-\frac{u'^2}{2\sigma^2}\right)}, \quad u \in \left[-\left\lfloor \frac{k}{2} \right\rfloor, \ldots, \left\lfloor \frac{k}{2} \right\rfloor\right]. \tag{2}$$

Utilizing the separable property of two-dimensional isotropic Gaussian functions $G(u, v) = g(u)\, g(v)$, we decompose a two-dimensional convolution into a cascade of two one-dimensional convolutions:

$$\hat{E} * G = (\hat{E} * g_x) * g_y. \tag{3}$$

Next, depthwise convolution is applied, enabling the same Gaussian kernel to be applied independently to each channel. Specifically, the smoothed embedding feature map $\tilde{E}$ is obtained by performing channel-wise convolutions sequentially along the horizontal and vertical directions Eqs. (4) and (5), and the convolutional weights W are fixed to g.

$$\tilde{E}^{(1)} = \mathrm{DWConv2d}\big(\tilde{E},\ W_x \in \mathbb{R}^{D\times 1\times 1\times k},\ \mathrm{pad} = (0, \lfloor k/2 \rfloor)\big), \tag{4}$$

$$\tilde{E}' = \mathrm{DWConv2d}\big(\tilde{E}^{(1)},\ W_y \in \mathbb{R}^{D\times 1\times k\times 1},\ \mathrm{pad} = (\lfloor k/2 \rfloor, 0)\big). \tag{5}$$

Adaptive Region-Aware Instantiation. To reasonably assign computational regions to each instance, we employ adaptive radius calculation based on the NNEC between points. This approach automatically infers the local supervision range for each point according to the spatial distribution of annotated points. First, we calculate the Euclidean distance matrix between all annotated

Fig. 5: Pseudo-Mask Selection. We utilize a low threshold to obtain pseudo points, thereby generating all possible masks and ensuring the reliability of background. Then, based on the probability, we select the mask associated with the highest probability as the final valid mask.

points in the sample. If there is exactly one annotated point, the radius degenerates into a global scale estimate. Radius for each point is defined as:

$$r_i = \begin{cases} \min\limits_{j \neq i} \mathbf{D}_{ij}, & N \geq 2 \\ \min(H, W) \cdot 0.5, & N = 1 \end{cases}, \quad \mathbf{D}_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\|_2. \tag{6}$$

We then define the spatial coordinate grid G, so the area contributing to the calculation for each final point is:

$$C_i(y, x) = \mathbf{1}\left[\|\mathbf{G}(y, x) - \mathbf{p}_i\|_2 \leq r_i\right], \quad \mathbf{G} \in \mathbb{R}^{H \times W \times 2}. \tag{7}$$

Therefore, for each point, the corresponding GT or pseudo-mask is used to segment positive and negative sample regions within the candidate area $\mathcal{R}$:

$$\mathcal{R}_i^+ = C_i \cap \{(y, x) \mid \mathbf{M}(y, x) = \ell_i\}, \quad \mathcal{R}_i^- = C_i \cap \{(y, x) \mid \mathbf{M}(y, x) \neq \ell_i\}, \tag{8}$$

where ℓ_i is the instance ID of point p_i.

Differentiable Center Sampling. Next, to obtain the embedding vector at each annotation point as the instance prototype, we use differentiable grid sampling, and $\bar{\mathbf{p}}_i$ is the normalized coordinate, placing the annotation point at the sub-pixel position. Thus, the formulation of the center feature is:

$$\mathbf{c}_i = \text{GridSample}(\tilde{\mathbf{E}}, \bar{\mathbf{p}}_i) \in \mathbb{R}^D, \quad \bar{\mathbf{p}}_i = \frac{2 \cdot \mathbf{p}_i}{(W - 1,\, H - 1)} - 1. \tag{9}$$

Consequently, the final calculation of maskhead discriminative loss is combined with Pull force Loss and Push force Loss:

$$\mathcal{L}_i = \frac{1}{|\mathcal{R}_i|} \sum_{(y,x) \in \mathcal{R}_i} \left[\alpha_{y,x} \left(\left\|\tilde{\mathbf{E}}(y, x) - \mathbf{c}_i\right\|_2 - (\tau - \alpha_{y,x} \cdot \delta)\right)\right]_+, \tag{10}$$

where, $[x]_+ = \max(x, 0)$ serves as the hinge function, and $\alpha_{y,x} \in \{+1, -1\}$ denotes the region label, with $+1$ for positive pixels and -1 for negative pixels. τ and δ are hyperparameters obtained from the experiment shown in supplement.

Instance Segmentation Generation During the inference stage and pseudo-mask generation, given the feature map $\mathbf{E} \in \mathbb{R}^{D\times H\times W}$ output by the trained embedding network and the predicted human location point prompts $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^{N}$, the overall generation of crowd segmentation is aligned with the aforementioned loss calculation. However, we observed that during the early stages of training, the generated pseudo-masks often exhibited incomplete clipping. To mitigate this issue, we propose incorporating additional spatial discriminative information by introducing a normalized squared geometric distance as a spatial regularization term. The computed embedding distance and geometric distance are fused into a joint energy function as follows:

$$E_i(y, x) = \left\| \tilde{\mathbf{E}}(y, x) - \mathbf{c}_i \right\|_2 + \lambda \cdot \frac{\|G(y, x) - p_i\|_2^2}{(r_i + \epsilon)^2}, \tag{11}$$

where λ is the geometry weight and $\tilde{\mathbf{E}}$ is the smoothed feature map mentioned above. Defining τ_g as an energy threshold in the inference, the final segmentation S(y,x) of the instance for the entire image is as follows:

$$S(y, x) = \begin{cases} i, & \text{if } E_i(y, x) < \tau_g \\ 0, & \text{otherwise} \end{cases}, \tag{12}$$

where i is the id number of each point.

3.3 Mask Constraint Counting

Given that region masks still exhibit inaccuracies and lack semantic information, we propose incorporating mask constraints, utilizing the generated instance mask as a pseudo-mask. As illustrated in Fig. 2, we introduce mask-constraint optimization as the last step to formulate the whole feedback framework.

To mitigate teacher uncertainty in pseudo-mask generation, we perform instance segmentation on points with confidence above a low threshold, while only masks exceeding a higher threshold are retained as valid for loss computation. This strategy improves the reliability of pseudo-mask learning, as visualized in Fig. 5. Based on this, we designed two computational methods as follows:

Background Penalty Loss. The purpose of this loss function is to ensure that no response is present in the background region. As described in Eq. (13), gradients are induced only when the prediction for true background pixels is positive, driving them toward zero or negative values, and those that are already negative do not contribute additional loss. Here, $\Omega_0 = \{p \mid M_{\text{gt}}(p) = 0\}$ denotes the background pixel geometry, and $\hat{f} \in \mathbb{R}^{H\times W}$ represents the bilinear interpolation output upscaled to the original image resolution.

$$L_{\text{bg}} = \frac{1}{|\Omega_0|} \sum_{p \in \Omega_0} \max\left(0, \hat{f}(p)\right). \tag{13}$$

Foreground Constraint Loss. To enforce instance-level consistency within each mask, we constrain both the aggregated foreground response and the number of activated pixels, with the one-point constraint serving as the primary objective. The foreground constraint loss is defined as:

$$\mathcal{L}_{\text{fg}} = \frac{1}{|K_v|} \sum_{k \in K_v} \left(\left| N_k^+ - 1 \right| + \lambda \left| S_k - 1 \right| \right), \tag{14}$$

$$S_k = \sum_{p \in \Omega_k} \max(0, \hat{f}(p)), \quad N_k^+ = \sum_{p \in \Omega_k} \mathbf{1}[\hat{f}(p) > 0], \tag{15}$$

where K_v denotes the set of valid instance IDs after pseudo-label filtering, and $\Omega_k = \{p \mid M_{\text{gt}}(p) = k\}$ represents the pixel set for the k-th instance.

Prior to this, we also attempted to use existing segmentation large models (such as SAM [21] and FastSAM [66]) as a method for generating pseudo-masks. However, their prohibitive inference cost in crowded scenes limits practicality.

4 Experiments

We evaluate the proposed method for crowd instance segmentation and counting on three crowd counting datasets: ShTech A [65], UCF-QNRF [18], and JHU++ [47]. Following P2R [30], DAC [27], and OT-M [29], three protocols are applied: 5%, 10%, and 40% of labeled data, with the remaining crowd images involved in training without annotations. The labeled samples in each percentage are chosen to align with DAC [27]. In ablation studies, we evaluated the impact of different architectures and hyperparameters to demonstrate the validity of our framework and the performance of our results.

Settings For all three datasets, the random crop with a size of 256×256 is implemented, and we limit the shorter side of each image to 1536 and 2048 for UCF-QNRF [18] and JHU++ [47], respectively while ShTech A [65] is still the original size. For data augmentation, labeled data is directly input into the student model, while unlabeled data is weakly augmented for the teacher model and strongly augmented for the student model.

4.1 Semi-supervised Crowd Instance Segmentation

This section demonstrates semi-supervised crowd instance segmentation performance in three datasets. As shown in Fig. 6, our XMask approach is compared with three different point-prompt methods. All methods utilize identical point predictions from a unified semi-supervised counting model, ensuring consistent localization quality and isolating the segmentation module's contribution. Our approach demonstrates excellent performance in crowd segmentation tasks,

Fig. 6: Visualization of masks generated by FastSAM-based, XMask, and ground truth generated by EDP-SAM. Through feature learning, our method demonstrates strong robustness in human segmentation, potentially segmenting other irrelevant objects compared to FastSAM. Furthermore, incorporating NNEC enhances the segmentation's dimensional adaptability.

achieving high recall masks across all test datasets. This is achieved through the utilization of predicted point constraints and the design of exclusivity feature learning methods.

As a baseline, FastSAM [66] reformulates SAM into a real-time framework to generate class-agnostic masks with prompt-guided refinement. The FastSAM-only variant assigns each point to the smallest candidate mask covering it. In contrast, FastSAM + NNEC introduces a conditional fallback strategy, using an NNEC-generated circular proxy when no valid mask is found.

As presented in Tab. 1, we evaluate IoU@0.5 with segmentation generated by EDP-SAM. Our XMask method consistently achieves excellent performance with substantial improvements over baselines. Ablation between our method variants reveals the complementary role of the geometric prior: on denser datasets (UCF-QNRF [18] and JHU++ [47]), the NNEC-augmented version yields consistent improvements, validating the conditional circle fallback for points lacking valid mask coverage. XMask also demonstrates significant computational efficiency, averaging 0.17s per image on ShTech A compared to 1.01s for FastSAM [66] (5.9× speedup). More efficiency comparisons are demonstrated in Sec. 4.3.

To further strengthen the reliability of results, we conduct comprehensive hyperparameter sensitivity analyses on XMask (λ and τ_g in Eqs. (11) and (12)) and FastSAM [66] (IoU and conf), which are shown in the supplementary material. Additionally, for possible circularity concern, we clarify that both baseline CrowdSAM [7] and FastSAM [66] employ the same NNEC-based constraint for fair comparison and the mask annotations were refined through manual correction (e.g., dense regions, boundary errors) to mitigate potential bias

4.2 Semi-Supervised Crowd Counting

We further evaluate the crowd counting accuracy against the previous state-of-the-art methods. Tab. 2 shows that our framework consistently outperforms across nearly all three benchmarks, with improvements particularly pronounced

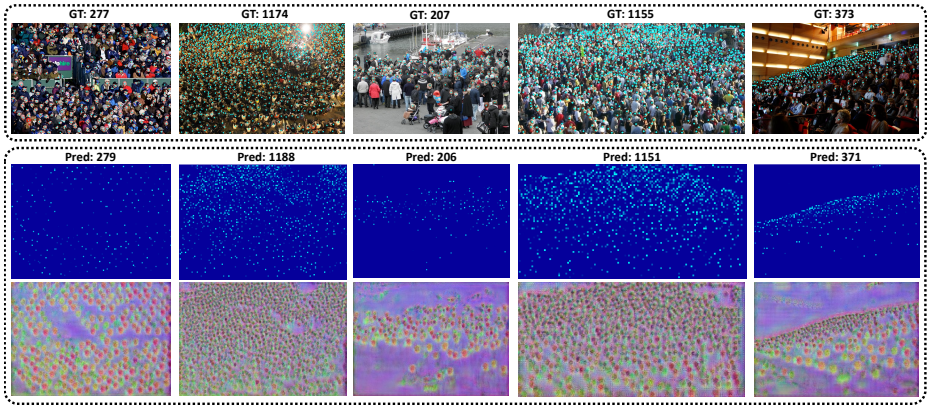

Fig. 7: Visualization of Counting results on ShTech A. The first row: GT location. The second row: predicted points. The third row: feature map from maskhead in our method. In both dense and normal crowd scenes, our method yields counting predictions that closely align with the ground-truth counts, while the learned feature maps maintain structural coherence with the semantic layout of crowd head regions.

Table 1: Semi-supervised crowd instance segmentation comparison on IoU@0.5.

Methods	NNEC	ShTech A [65]			UCF-QNRF [18]			JHU++ [47]		
		5%	10%	40%	5%	10%	40%	5%	10%	40%
FastSAM [66]		0.279	0.281	0.291	0.203	0.211	0.213	0.186	0.185	0.189
FastSAM [66]	✓	0.287	0.290	0.300	0.223	0.229	0.233	0.197	0.182	0.201
S4M [61]		0.005	0.007	0.011	0.010	0.012	0.015	0.013	0.014	0.021
CrowdSAM [7]	✓	0.201	0.201	0.212	0.149	0.156	0.163	0.187	0.189	0.192
point only	✓	0.195	0.194	0.202	0.198	0.201	0.207	0.180	0.182	0.183
XMask(ours)		**0.314**	**0.318**	0.331	0.269	0.277	0.278	0.249	0.253	0.257
XMask (ours)	✓	0.313	**0.318**	**0.333**	**0.282**	**0.290**	**0.297**	**0.258**	**0.263**	**0.270**

under limited annotated images. By leveraging reliable pseudo-masks and structural consistency learning from XMask, our method outperforms P2R [30], the base counting model used in Stage 1. As visualized in Fig. 7, our approach delivers accurate counting predictions while enforcing structural coherence between features and head-region semantics.

Under the 5% setting, our proposed framework achieves the best performance in nearly all protocols across all crowd datasets, demonstrating that the learned segmentation priors effectively compensate for limited supervision. As label ratios increase to 10% and 40%, our method maintains consistent gains, with notably substantial MSE improvements, indicating enhanced robustness against large estimation outliers.

4.3 Ablation study

We then ablate the components in the proposed method and analyze the hyperparameters.

Table 2: Semi-supervised crowd counting performance comparison between recent methods and our framework on three datasets under various labeled protocols.

labeled Pct.	Methods	ShTech A [65]		UCF-QNRF [18]		JHU++ [47]	
		MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓
5%	MT [50]	104.7	33.2	172.4	284.9	101.5	363.5
	L2R [34]	103.0	27.6	160.1	272.3	101.4	338.8
	DAC [27]	85.4	134.5	120.2	209.3	82.2	294.9
	OT-M [29]	83.7	133.3	118.4	195.4	82.7	304.5
	P2R [30]	69.9	119.5	100.1	182.5	77.8	293.5
	Ours	**64.6**	**106.2**	**98.4**	**163.7**	**77.1**	**289.5**
10%	MT [50]	319.3	94.5	156.1	245.5	250.3	90.2
	L2R [34]	90.3	153.5	148.9	249.8	87.5	315.3
	DAC [27]	74.9	115.5	109.0	187.2	75.9	282.3
	OT-M [29]	80.1	118.5	113.1	186.7	73.0	280.6
	P2R [30]	64.2	114.6	94.9	167.2	**68.7**	272.3
	Ours	**62.4**	**104.4**	**93.9**	**155.7**	**68.7**	**258.7**
40%	MT [50]	88.2	151.1	147.2	249.6	121.5	388.9
	L2R [34]	86.5	148.2	145.1	256.1	123.6	376.1
	DAC [27]	67.5	110.7	91.1	153.4	65.1	260.0
	OT-M [29]	70.7	114.5	100.6	167.6	72.1	272.0
	P2R [30]	**55.6**	95.0	86.0	144.3	63.3	271.1
	Ours	56.2	**93.5**	**84.0**	**143.6**	**61.9**	**256.1**

Segmentation Efficiency. We conducted segmentation inference experiments on the ShTechA test dataset, which was divided into multiple density intervals for comparison. As shown in Fig. 8, compared to FastSAM and SAM, our method significantly improves segmentation speed, achieving less than 0.3 seconds per image across all density intervals, while demonstrating insensitivity to density. Therefore, our approach is highly advantageous for training efficiency when generating pseudo masks in semi-supervised settings.

Loss Function Comparison. In semi-supervised segmentation learning, we propose utilizing the maskhead discriminative loss. We compared it with several other segmentation losses on the ShTechA dataset with 5% labeled data, including Dice loss and Focal loss. Results in Tab. 3a demonstrate that our approach achieves the best performance in terms of mask IoU and F1 score.

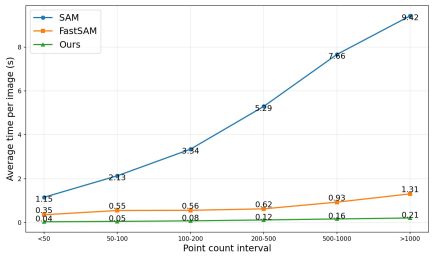

Fig. 8: Our method consistently outperforms SAM and FastSAM in segmentation efficiency across varying crowd densities, maintaining shorter processing times despite differences in density.

Differentiable Center Sampling. Experiments on ShTech A (Tab. 3c) show that feature sampling slightly improves performance. Despite the modest gain, we adopt differentiable sampling over discrete extraction due to its sub-pixel precision and smoother gradient propagation, which lead to more stable and optimization-friendly feature embeddings.

Depthwise Gaussian Smoothing. This module applies Gaussian smoothing to feature embeddings before pairwise L2 distance computation, enhancing

Table 3: Comprehensive ablation studies on different components and settings.

(a) Loss function comparison on ShTech A. Dice loss calculates overlap-based ratio between prediction and ground truth, while Focal loss weighted cross-entropy focusing on hard samples.

Loss	IoU↑	F1↑
Dice loss	0.2381	0.0946
Focal loss	0.2893	0.2070
Ours	**0.3119**	**0.2574**

(b) Gaussian Smoothing on ShTech A

Gaussian	K	σ	IoU↑	F1↑
	/	/	0.3050	0.2481
✓	7	1.5	0.3119	0.2574
✓	5	1.5	0.3111	0.2566
✓	9	1.5	0.3120	0.2577
✓	7	3	**0.3130**	**0.2584**
✓	5	3	0.3119	0.2575
✓	9	3	0.3130	0.2570

(c) Differentiable Center Sampling on UCF

Sample	5%		10%		40%	
	IoU↑	F1↑	IoU↑	F1↑	IoU↑	F1↑
	0.2806	0.1973	0.2888	0.2111	0.2955	0.2079
✓	**0.2807**	**0.1976**	**0.2892**	**0.2129**	**0.2956**	**0.2079**

(d) Optimization Modules on ShTech A

Head	Backbone	Params	MAE↓	MSE↓
✓	✓	16.77M	68.1	106.8
✓		15.58M	70.2	111.3
	✓	2.05M	**64.6**	**106.2**

spatial coherence and reducing noisy activations. Tab. 3b shows that Gaussian filtering consistently improves segmentation performance and a kernel size of 7 with $\sigma = 3$ achieves the best results.

Trained Modules. In the final counting stage, we explore different fine-tuning strategies to balance performance and efficiency. As shown in Tab. 3d, unfreezing only the decoder achieves the best counting accuracy. Therefore, we adopt this setting as the default configuration.

5 Conclusion

In this work, we propose to supervise the semi-supervised crowd instance segmentation and counting based on mask supervision, which provides more structural individual shape information than point annotation. First, we propose an Exclusion-Constrained Dual-Prompt SAM (EDP-SAM) to generate robust masks from point annotations. Next, we combine EDP-SAM with manual correction to generate accurate mask annotations for the existing datasets. To better segment instances in crowded images, we propose Exclusivity-Guided Mask Learning (XMask) that enforces spatial separation through a discriminative mask objective. Extensive experiments demonstrate that mask supervision effectively improves semi-supervised crowd instance segmentation and counting performance, offering a scalable solution to annotation scarcity. Future work will focus on improving boundary precision and refining point-generation strategies to achieve finer-grained segmentation quality.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Project 62406090 and Jiangsu Science and Technology Major Program (Grant No. BG2024041).

References

1. Abousamra, S., Hoai, M., Samaras, D., Chen, C.: Localization in the crowd with topological constraints. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. pp. 872–881. AAAI Press (2021). https://doi.org/10.1609/AAAI.V35I2.16170, https://doi.org/10.1609/aaai.v35i2.16170
2. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2282 (2012). https://doi.org/10.1109/TPAMI.2012.120, https://doi.org/10.1109/TPAMI.2012.120
3. Bai, M., Urtasun, R.: Deep watershed transform for instance segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 2858–2866. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.305, https://doi.org/10.1109/CVPR.2017.305
4. Berrada, T., Couprie, C., Alahari, K., Verbeek, J.: Guided distillation for semi-supervised instance segmentation. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024. pp. 464–472. IEEE (2024). https://doi.org/10.1109/WACV57701.2024.00053, https://doi.org/10.1109/WACV57701.2024.00053
5. Brabandere, B.D., Neven, D., Gool, L.V.: Semantic instance segmentation with a discriminative loss function. CoRR **abs/1708.02551** (2017), http://arxiv.org/abs/1708.02551
6. Cai, Z., Ravichandran, A., Maji, S., Fowlkes, C.C., Tu, Z., Soatto, S.: Exponential moving average normalization for self-supervised and semi-supervised learning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 194–203. Computer Vision Foundation / IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.00026, https://openaccess.thecvf.com/content/CVPR2021/html/Cai_Exponential_Moving_Average_Normalization_for_Self-Supervised_and_Semi-Supervised_Learning_CVPR_2021_paper.html
7. Cai, Z., Gao, Y., Zheng, Y., Zhou, N., Huang, D.: Crowd-sam: SAM as a smart annotator for object detection in crowded scenes. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXIX. Lecture Notes in Computer Science, vol. 15127, pp. 334–351. Springer (2024). https://doi.org/10.1007/978-3-031-72890-7_20, https://doi.org/10.1007/978-3-031-72890-7_20
8. Chen, J., Wang, Z.: Multi-task semi-supervised crowd counting via global to local self-correction. Pattern Recognit. **140**, 109506 (2023). https://doi.org/10.1016/J.PATCOG.2023.109506, https://doi.org/10.1016/j.patcog.2023.109506
9. Chen, Y., Wang, Q., Yang, J., Chen, B., Xiong, H., Du, S.: Learning discriminative features for crowd counting. IEEE Trans. Image Process. **33**, 3749–3764 (2024). https://doi.org/10.1109/TIP.2024.3408609, https://doi.org/10.1109/TIP.2024.3408609
10. Cheng, T., Wang, X., Huang, L., Liu, W.: Boundary-preserving mask R-CNN. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020

- 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV. Lecture Notes in Computer Science, vol. 12359, pp. 660–676. Springer (2020). https://doi.org/10.1007/978-3-030-58568-6_39, https://doi.org/10.1007/978-3-030-58568-6_39

11. Chu, X., Zheng, A., Zhang, X., Sun, J.: Detection in crowded scenes: One proposal, multiple predictions. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 12211–12220. Computer Vision Foundation / IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.01223, https://openaccess.thecvf.com/content_CVPR_2020/html/Chu_Detection_in_Crowded_Scenes_One_Proposal_Multiple_Predictions_CVPR_2020_paper.html
12. Ding, H., Qiao, S., Yuille, A.L., Shen, W.: Deeply shape-guided cascade for instance segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 8278–8288. Computer Vision Foundation / IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.00818, https://openaccess.thecvf.com/content/CVPR2021/html/Ding_Deeply_Shape-Guided_Cascade_for_Instance_Segmentation_CVPR_2021_paper.html
13. Filipiak, D., Zapala, A., Tempczyk, P., Fensel, A., Cygan, M.: Polite teacher: Semi-supervised instance segmentation with mutual learning and pseudo-label thresholding. IEEE Access **12**, 37744–37756 (2024). https://doi.org/10.1109/ACCESS.2024.3374073, https://doi.org/10.1109/ACCESS.2024.3374073
14. Ge, Z., Jie, Z., Huang, X., Xu, R., Yoshie, O.: PS-RCNN: detecting secondary human instances in a crowd via primary object suppression. In: IEEE International Conference on Multimedia and Expo, ICME 2020, London, UK, July 6-10, 2020. pp. 1–6. IEEE (2020). https://doi.org/10.1109/ICME46284.2020.9102793, https://doi.org/10.1109/ICME46284.2020.9102793
15. Gu, Z., Chen, H., Xu, Z.: Diffusioninst: Diffusion model for instance segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024. pp. 2730–2734. IEEE (2024). https://doi.org/10.1109/ICASSP48485.2024.10447191, https://doi.org/10.1109/ICASSP48485.2024.10447191
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. IEEE Trans. Pattern Anal. Mach. Intell. **42**(2), 386–397 (2020). https://doi.org/10.1109/TPAMI.2018.2844175, https://doi.org/10.1109/TPAMI.2018.2844175
17. Hu, J., Chen, C., Cao, L., Zhang, S., Shu, A., Jiang, G., Ji, R.: Pseudo-label alignment for semi-supervised instance segmentation. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 16291–16301. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.01497, https://doi.org/10.1109/ICCV51070.2023.01497
18. Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Máadeed, S., Rajpoot, N.M., Shah, M.: Composition loss for counting, density map estimation and localization in dense crowds. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II. Lecture Notes in Computer Science, vol. 11206, pp. 544–559. Springer (2018). https://doi.org/10.1007/978-3-030-01216-8_33, https://doi.org/10.1007/978-3-030-01216-8_33
19. Kang, K., Wang, X.: Fully convolutional neural networks for crowd segmentation. CoRR **abs/1411.4464** (2014), http://arxiv.org/abs/1411.4464
20. Ke, L., Danelljan, M., Li, X., Tai, Y., Tang, C., Yu, F.: Mask transfiner for high-quality instance segmentation. In: IEEE/CVF Conference on Computer Vision

and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 4402–4411. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.00437, https://doi.org/10.1109/CVPR52688.2022.00437

21. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B.: Segment anything. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 3992–4003. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.00371, https://doi.org/10.1109/ICCV51070.2023.00371
22. Lalit, M., Tomancak, P., Jug, F.: Embedseg: Embedding-based instance segmentation for biomedical microscopy data. Medical Image Anal. **81**, 102523 (2022). https://doi.org/10.1016/J.MEDIA.2022.102523, https://doi.org/10.1016/j.media.2022.102523
23. Lee, D.H.: Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks (2013), https://api.semanticscholar.org/CorpusID:18507866
24. Li, C., Hu, X., Abousamra, S., Chen, C.: Calibrating uncertainty for semi-supervised crowd counting. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 16685–16695. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.01534, https://doi.org/10.1109/ICCV51070.2023.01534
25. Li, Y., Zhang, X., Chen, D.: Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 1091–1100. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00120, http://openaccess.thecvf.com/content_cvpr_2018/html/Li_CSRNet_Dilated_Convolutional_CVPR_2018_paper.html
26. Liang, Z., Li, Z., Xu, S., Tan, M., Jia, K.: Instance segmentation in 3d scenes using semantic superpoint tree networks. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 2763–2772. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00278, https://doi.org/10.1109/ICCV48922.2021.00278
27. Lin, H., Ma, Z., Hong, X., Wang, Y., Su, Z.: Semi-supervised crowd counting via density agency. In: Magalhães, J., Bimbo, A.D., Satoh, S., Sebe, N., Alameda-Pineda, X., Jin, Q., Oria, V., Toni, L. (eds.) MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022. pp. 1416–1426. ACM (2022). https://doi.org/10.1145/3503161.3547867, https://doi.org/10.1145/3503161.3547867
28. Lin, H., Ma, Z., Ji, R., Wang, Y., Su, Z., Hong, X., Meng, D.: Semi-supervised counting via pixel-by-pixel density distribution modelling. CoRR **abs/2402.15297** (2024). https://doi.org/10.48550/ARXIV.2402.15297, https://doi.org/10.48550/arXiv.2402.15297
29. Lin, W., Chan, A.B.: Optimal transport minimization: Crowd localization on density maps for semi-supervised counting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 21663–21673. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.02075, https://doi.org/10.1109/CVPR52729.2023.02075
30. Lin, W., Zhao, C., Chan, A.B.: Point-to-region loss for semi-supervised point-based crowd counting. In: IEEE/CVF Conference on Computer Vision and Pat-

tern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 29363–29373. Computer Vision Foundation / IEEE (2025). https://doi.org/10.1109/CVPR52734.2025.02734, https://openaccess.thecvf.com/content/CVPR2025/html/Lin_Point-to-Region_Loss_for_Semi-Supervised_Point-Based_Crowd_Counting_CVPR_2025_paper.html

31. Liu, S., Jia, J., Fidler, S., Urtasun, R.: SGN: sequential grouping networks for instance segmentation. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 3516–3524. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.378, https://doi.org/10.1109/ICCV.2017.378
32. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 8759–8768. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00913, http://openaccess.thecvf.com/content_cvpr_2018/html/Liu_Path_Aggregation_Network_CVPR_2018_paper.html
33. Liu, S., Huang, D., Wang, Y.: Adaptive NMS: refining pedestrian detection in a crowd. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 6459–6468. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00662, http://openaccess.thecvf.com/content_CVPR_2019/html/Liu_Adaptive_NMS_Refining_Pedestrian_Detection_in_a_Crowd_CVPR_2019_paper.html
34. Liu, X., van de Weijer, J., Bagdanov, A.D.: Leveraging unlabeled data for crowd counting by learning to rank. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 7661–7669. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00799, http://openaccess.thecvf.com/content_cvpr_2018/html/Liu_Leveraging_Unlabeled_Data_CVPR_2018_paper.html
35. Liu, X., van de Weijer, J., Bagdanov, A.D.: Exploiting unlabeled data in cnns by self-supervised learning to rank. IEEE Trans. Pattern Anal. Mach. Intell. **41**(8), 1862–1878 (2019). https://doi.org/10.1109/TPAMI.2019.2899857, https://doi.org/10.1109/TPAMI.2019.2899857
36. Liu, Y., Liu, L., Wang, P., Zhang, P., Lei, Y.: Semi-supervised crowd counting via self-training on surrogate tasks. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XV. Lecture Notes in Computer Science, vol. 12360, pp. 242–259. Springer (2020). https://doi.org/10.1007/978-3-030-58555-6_15, https://doi.org/10.1007/978-3-030-58555-6_15
37. Liu, Y., Shi, M., Zhao, Q., Wang, X.: Point in, box out: Beyond counting persons in crowds. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 6469–6478. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00663, http://openaccess.thecvf.com/content_CVPR_2019/html/Liu_Point_in_Box_Out_Beyond_Counting_Persons_in_Crowds_CVPR_2019_paper.html
38. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1) (Jan 2024). https://doi.org/10.1038/s41467-024-44824-z, http://dx.doi.org/10.1038/s41467-024-44824-z
39. Ma, Z., Wei, X., Hong, X., Gong, Y.: Bayesian loss for crowd count estimation with point supervision. In: 2019 IEEE/CVF International Conference on Computer

Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 6141–6150. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00624, https://doi.org/10.1109/ICCV.2019.00624

40. Meng, Y., Zhang, H., Zhao, Y., Yang, X., Qian, X., Huang, X., Zheng, Y.: Spatial uncertainty-aware semi-supervised crowd counting. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 15529–15539. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.01526, https://doi.org/10.1109/ICCV48922.2021.01526
41. Molina, J.M., Llerena, J.P., Usero, L., Patricio, M.A.: Advances in instance segmentation: Technologies, metrics and applications in computer vision. Neurocomputing **625**, 129584 (2025). https://doi.org/10.1016/J.NEUCOM.2025.129584, https://doi.org/10.1016/j.neucom.2025.129584
42. Montalvo, J., García-Martín, Á., Carballeira, P., SanMiguel, J.C.: Unsupervised class generation to expand semantic segmentation datasets. J. Imaging **11**(6), 172 (2025). https://doi.org/10.3390/JIMAGING11060172, https://doi.org/10.3390/jimaging11060172
43. Morales-Brotons, D., Vogels, T., Hendrikx, H.: Exponential moving average of weights in deep learning: Dynamics and benefits. Trans. Mach. Learn. Res. **2024** (2024), https://openreview.net/forum?id=2M9CUnYnBA
44. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031, https://doi.org/10.1109/TPAMI.2016.2577031
45. Sam, D.B., Peri, S.V., Sundararaman, M.N., Kamath, A., Babu, R.V.: Locate, size, and count: Accurately resolving people in dense crowds via detection. IEEE Trans. Pattern Anal. Mach. Intell. **43**(8), 2739–2751 (2021). https://doi.org/10.1109/TPAMI.2020.2974830, https://doi.org/10.1109/TPAMI.2020.2974830
46. Shang, C., Wu, Q., Meng, F., Xu, L.: Instance segmentation by learning deep feature in embedding space. In: 2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019. pp. 2444–2448. IEEE (2019). https://doi.org/10.1109/ICIP.2019.8803021, https://doi.org/10.1109/ICIP.2019.8803021
47. Sindagi, V.A., Yasarla, R., Patel, V.M.: JHU-CROWD++: large-scale crowd counting dataset and A benchmark method. IEEE Trans. Pattern Anal. Mach. Intell. **44**(5), 2594–2609 (2022). https://doi.org/10.1109/TPAMI.2020.3035969, https://doi.org/10.1109/TPAMI.2020.3035969
48. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C., Cubuk, E.D., Kurakin, A., Li, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html
49. Song, Q., Wang, C., Jiang, Z., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Wu, Y.: Rethinking counting and localization in crowds: A purely point-based framework. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 3345–3354. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00335, https://doi.org/10.1109/ICCV48922.2021.00335

50. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A.: Striving for simplicity: The all convolutional net. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings (2015), http://arxiv.org/abs/1412.6806
51. Stewart, R., Andriluka, M., Ng, A.Y.: End-to-end people detection in crowded scenes. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 2325–2333. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.255, https://doi.org/10.1109/CVPR.2016.255
52. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 1195–1204 (2017), https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html
53. Ulmer, M., Boerdijk, W., Triebel, R., Durner, M.: Conditional latent diffusion models for zero-shot instance segmentation. In: IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025. pp. 24360–24369. IEEE (2025). https://doi.org/10.1109/ICCV51701.2025.02258, https://doi.org/10.1109/ICCV51701.2025.02258
54. Wang, D., Zhang, J., Du, B., Xu, M., Liu, L., Tao, D., Zhang, L.: SAMRS: scaling-up remote sensing segmentation dataset with segment anything model. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/1be3843e534ee06d3a70c7f62b983b31-Abstract-Datasets_and_Benchmarks.html
55. Wang, X., Zhan, Y., Zhao, Y., Yang, T., Ruan, Q.: Semi-supervised crowd counting with spatial temporal consistency and pseudo-label filter. IEEE Trans. Circuits Syst. Video Technol. **33**(8), 4190–4203 (2023). https://doi.org/10.1109/TCSVT.2023.3241175, https://doi.org/10.1109/TCSVT.2023.3241175
56. Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C.: Repulsion loss: Detecting pedestrians in a crowd. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 7774–7783. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00811, http://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Repulsion_Loss_Detecting_CVPR_2018_paper.html
57. Wang, Z., Li, Y., Wang, S.: Noisy boundaries: Lemon or lemonade for semi-supervised instance segmentation? In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 16805–16814. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01632, https://doi.org/10.1109/CVPR52688.2022.01632
58. Wang, Z., Li, Y., Wang, S.: Noisy boundaries: Lemon or lemonade for semi-supervised instance segmentation? In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 16805–16814. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01632, https://doi.org/10.1109/CVPR52688.2022.01632

59. Xun, Z., Ku, Y.: Multi-expert diffusion segmentation for semi-supervised 3d medical image segmentation (2026), https://doi.org/10.20944/preprints202601.0942.v1
60. Ying, H., Huang, Z., Liu, S., Shao, T., Zhou, K.: Embedmask: Embedding coupling for instance segmentation. In: Zhou, Z.H. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. pp. 1266–1273. International Joint Conferences on Artificial Intelligence Organization (8 2021). https://doi.org/10.24963/ijcai.2021/175, https://doi.org/10.24963/ijcai.2021/175, main Track
61. Yoon, H., Shin, H., Hong, E., Choi, H., Cho, H., Jeong, D., Kim, S.: S^4m: Boosting semi-supervised instance segmentation with SAM. In: IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025. pp. 20226–20236. IEEE (2025). https://doi.org/10.1109/ICCV51701.2025.01881, https://doi.org/10.1109/ICCV51701.2025.01881
62. Zhang, G., Lu, X., Tan, J., Li, J., Zhang, Z., Li, Q., Hu, X.: Refinemask: Towards high-quality instance segmentation with fine-grained features. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 6861–6869. Computer Vision Foundation / IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.00679, https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_RefineMask_Towards_High-Quality_Instance_Segmentation_With_Fine-Grained_Features_CVPR_2021_paper.html
63. Zhang, S., Ke, W., Liu, S., Hong, X., Zhang, T.: Boosting semi-supervised crowd counting with scale-based active learning. In: Cai, J., Kankanhalli, M.S., Prabhakaran, B., Boll, S., Subramanian, R., Zheng, L., Singh, V.K., César, P., Xie, L., Xu, D. (eds.) Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024. pp. 8681–8690. ACM (2024). https://doi.org/10.1145/3664647.3680976, https://doi.org/10.1145/3664647.3680976
64. Zhang, Y., Lv, B., Xue, L., Zhang, W., Liu, Y., Fu, Y., Cheng, Y., Qi, Y.: Semisam+: Rethinking semi-supervised medical image segmentation in the era of foundation models. Medical Image Anal. **106**, 103733 (2025). https://doi.org/10.1016/J.MEDIA.2025.103733, https://doi.org/10.1016/j.media.2025.103733
65. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 589–597. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.70, https://doi.org/10.1109/CVPR.2016.70
66. Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., Wang, J.: Fast segment anything. CoRR **abs/2306.12156** (2023). https://doi.org/10.48550/ARXIV.2306.12156, https://doi.org/10.48550/arXiv.2306.12156
67. Zou, Y., Liu, Z., Gu, Y., Du, B., Xu, Y.: Consistent-point: Consistent pseudo-points for semi-supervised crowd counting and localization. CoRR **abs/2503.12441** (2025). https://doi.org/10.48550/ARXIV.2503.12441, https://doi.org/10.48550/arXiv.2503.12441