7-2023

Reducing Spatial Labeling Redundancy for Active Semi-Supervised Crowd Counting

Yongtuo LIU

Sucheng REN

Liangyu CHAI

Hanjie WU

Dan XU

See next page for additional authors

Short Paper

Reducing Spatial Labeling Redundancy for Active Semi-Supervised Crowd Counting

Yongtuo Liu, Sucheng Ren, Liangyu Chai, Hanjie Wu, Dan Xu, Jing Qin, *Senior Member, IEEE*, and Shengfeng He, *Senior Member, IEEE*

***Abstract*—Labeling is onerous for crowd counting as it should annotate each individual in crowd images. Recently, several methods have been proposed for semi-supervised crowd counting to reduce the labeling efforts. Given a limited labeling budget, they typically select a few crowd images and densely label all individuals in each of them. Despite the promising results, we argue the None-or-All labeling strategy is suboptimal as the densely labeled individuals in each crowd image usually appear similar while the massive unlabeled crowd images may contain entirely diverse individuals. To this end, we propose to break the labeling chain of previous methods and make the first attempt to reduce spatial labeling redundancy for semi-supervised crowd counting. First, instead of annotating all the regions in each crowd image, we propose to annotate the representative ones only. We analyze the region representativeness from both vertical and horizontal directions of initially estimated density maps, and formulate them as cluster centers of Gaussian Mixture Models. Additionally, to leverage the rich unlabeled regions, we exploit the similarities among individuals in each crowd image to directly supervise the unlabeled regions via feature propagation instead of the error-prone label propagation employed in the previous methods. In this way, we can transfer the original spatial labeling redundancy caused by individual similarities to effective supervision signals on the unlabeled regions. Extensive experiments on the widely-used benchmarks demonstrate that our method can outperform previous best approaches by a large margin.**

***Index Terms*—Crowd counting, semi-supervised learning, spatial labeling redundancy.**

Manuscript received 8 July 2021; revised 31 October 2022; accepted 18 December 2022. Date of publication 28 December 2022; date of current version 5 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61972162; in part by Guangdong International Science and Technology Cooperation Project under Grant 2021A0505030009; in part by Guangdong Natural Science Foundation under Grant 2021A1515012625; in part by Guangzhou Basic and Applied Research Project under Grant 202102021074; in part by CCF-Tencent Open Research fund under Grant RAGR20210114; in part by the Project of Strategic Importance scheme of The Hong Kong Polytechnic University under Grant 1-ZE2Q; and in part by the Innovation and Technology Fund of Hong Kong Innovation and Technology Commission under Grant ITS/180/20FP. Recommended for acceptance by K. G. Derpanis. *(Corresponding author: Shengfeng He.)*

Yongtuo Liu, Sucheng Ren, Liangyu Chai, and Hanjie Wu are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: csmanlyt@mail.scut.edu.cn; oliverrensu@gmail.com; icepoint1018@gmail.com; cshanjiewu@gmail.com).

Dan Xu is with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong (e-mail: danxu@cse.ust.hk).

Jing Qin is with the Department of Nursing, Hong Kong Polytechnic University, Hung Hom, Hong Kong (e-mail: harry.qin@polyu.edu.hk).

Shengfeng He is with the School of Computing and Information Systems, Singapore Management University, Singapore 188065 (e-mail: shengfenghe7@gmail.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2022.3232712, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2022.3232712

I. Introduction

Crowd counting has drawn increasing attention in the community due to its essential role in social management, such as crowd monitoring and crowd congestion warning [1], [3]. Benefiting from the powerful CNN architectures, lots of works have been proposed and advanced the performance of crowd counting. Most of them are mainly dedicated to solving various challenges of crowd counting in a fully-supervised manner [4], [5], [6], [7], [8]. However, labeling for crowd counting is quite burdensome as we have to annotate each individual in crowd images.

To reduce the labeling efforts, we study crowd counting in a semi-supervised setting where only a small labeling budget is available. Methods in this line can be mainly grouped into three categories: (i) [9], [10], [11] leverage self-supervised constraints to learn a generic feature extractor from unlabeled crowd images. (ii) [12], [13] introduce knowledge transfer to bridge the labeled and unlabeled data. (iii) [14], [15], [16] exploit temporal labeling redundancy in crowd video scenarios.

Notwithstanding the demonstrated success of the above methods, they all view each crowd image as a minimum labeling unit and densely label all individuals in a limited number of crowd images. The None-or-All labeling strategy is suboptimal considering the labeling burden and labeling redundancy in each crowd image. (i) Compare to other computer vision tasks, the labeling burden of crowd counting mainly resides in each crowd image where hundreds of individuals may need to be annotated. The existing methods try to alleviate the labeling burden by decreasing the number of labeled crowd images, which seems palliative for the crowd counting problem. (ii) Individuals in each annotated crowd image usually appear similar with lots of labeling redundancy as they are captured in the same crowd scene. This makes the annotated individuals lack diversity and cannot adapt to various crowd scenes, e.g., different camera perspectives, weather and illumination conditions.

To this end, we propose to break the labeling chain of previous methods and make the first attempt to reduce spatial labeling redundancy for semi-supervised crowd counting. First, instead of annotating all the regions in each crowd image, we propose to annotate the representative ones only (see Fig. 1). We analyze the region representativeness from both the vertical and horizontal directions of initially estimated density maps, and design a Multi-level Density-aware Cluster (MDC) Strategy to formulate the representative regions as cluster centers of Gaussian Mixture Models based on their multi-level density vectors. In this way, our method can effectively reduce the spatial labeling redundancy in each crowd image and label more crowd images with various crowd scenes given the same labeling budget. Additionally, to leverage the rich unlabeled regions, we further exploit the similarities among individuals in each crowd image to directly supervise the unlabeled regions via feature propagation in a Crowd Affinity Propagation (CAP) module.

Fig. 1. Given a limited labeling budget (e.g., 10% of the entire dataset), all the previous methods adopt a None-or-All labeling strategy and select a few crowd images to densely label all the individuals which typically appear similar and lack diversity. Differently, we propose to break the labeling chain of previous methods and annotate the representative regions only in each crowd image.

The CAP module propagates crowd features of the unlabeled regions to update those of the labeled regions based on the feature affinities in the forward propagation. Then in the backward propagation, the unlabeled regions can be directly supervised by the labels of the labeled regions via feature backpropagation. After training, we can optionally remove the CAP module without performance degradation which makes it computationally free at the inference stage. By the CAP module, we can transfer the original spatial labeling redundancy caused by the individual similarities to effective supervision signals and directly supervise the unlabeled regions without generating the error-prone pseudo labels. Extensive experiments on widely-used benchmarks demonstrate our method outperforms previous approaches by a large margin. For example, our method outperforms the best AL-AC [13] by 9.4%/8.1% and 8.6%/22.5% for MAE/RMSE in the ShanghaiTec PartA and PartB datasets, respectively.

The contributions are summarized as follows: 1) We propose to break the labeling chain of previous methods and reduce the spatial labeling redundancy by annotating representative regions only for effective semi-supervised crowd counting. 2) We analyze the region representativeness from both vertical and horizontal directions and formulate representative regions as cluster centers of Gaussian Mixture Models. Furthermore, to leverage the unlabeled regions, we propose to exploit the similarities among individuals to directly supervise the unlabeled regions via feature propagation without the error-prone pseudo label generation. 3) Extensive experiments show that our method can achieve state-of-the-art performance and outperform previous best approaches by a large margin.

II. Related Work

A. Crowd Counting

Early methods for crowd counting are based on hand-crafted features (e.g., SIFT, Fourier Analysis, and HOG). They estimate crowd counts by either direct regression [17], [18], [19] or human parts detection [20], [21], [22]. Recently, a lot of CNN-based methods have been proposed and advanced the performance of crowd counting. Most of them mainly solve various challenges of crowd counting in a *fully-supervised* manner, including large scale variations [4], [5], [23], [24], [25], [26], attentive feature extraction [28], [29], [30], [31], [32], label noises [6], [7], empirical Gaussian kernel [33], [34], [35], estimation uncertainty [36], [37], structural constraints [8], [38], and etc. These methods require a great number of labeled data in the training process which are rather burdensome for crowd counting.

B. Semi-Supervised Crowd Counting

Recently, several methods are designed to learn a crowd counter with a limited labeling budget. They can be mainly grouped into three categories as follows:

Self-supervised Constraints [9], [10], [11]: Liu et al. [9] propose to exploit unlabeled data by ranking cropped patches according to their containment relationships. Sam et al. [10] extract useful feature representations by learning a Grid Winner-Take-ALL (GWTA) autoencoder from unlabeled crowd images. Liu et al. [11] propose to leverage surrogate tasks with IRAST constraints to train a generic feature extractor.

Knowledge Transfer [12], [13]: Sindagi et al. [12] introduce a Gaussian Process (GP) to generate pseudo labels of the unlabeled data. Zhao et al. [13] propose to transfer feature representations across labeled and unlabeled data by a distribution classifier with the mixup technique.

Temporal Redundancy [14], [15], [16]: Tan et al. [14] propose a Semi-Supervised Elastic Net (SSEN) to regularize temporally neighboring samples. Loy et al. [15] analyze the geometric structure of crowd patterns and design the distribution and temporal regularization for manifold learning. Zhou et al. [16] propose a submodular method to annotate informative frames in crowd videos and introduce the graph Laplacian regularization for semi-supervised learning.

Despite the promising results of the above methods, they all adopt a None-or-All labeling strategy which inevitably introduces lots of labeling redundancy and lack diversity. Differently, we propose to annotate the representative regions only in each crowd image and transfer the labeling redundancy caused by individual similarities to effective supervision signals on the unlabeled regions.

C. Weakly-Supervised Crowd Counting

There exist two kinds of image-level labels (i.e., total number of humans [39], [40], and density levels [28]) that are explored in weakly-supervised counting methods.

For the number labels [39], [40], ideally, we only need numbers of humans to implement weakly-supervised counting. However, in practice, we still need to count all the humans one by one to get the final number in the labeling process, which takes nearly the same effort as annotating all the head positions in the fully supervised setting. For the density labels [28], they cannot achieve counting by density labels only. This is because density labels cannot directly supervise the network to output an accurate number for counting. For example, the label of "low density" is undefined numerically. In practice, [28] relies on the fully annotated human heads in the source domain to achieve counting in the target domain by additionally labeling density labels. Besides, the boundaries between different density labels are ambiguous. For example, how to distinguish "very low density" and "low density" is still an open problem. Based on the analysis, semi-supervised setting considered in this paper plays a vital role in reducing the labeling effort while achieving promising counting performance.

III. Method

A. Framework Overview

As shown in Fig. 2, we propose a novel semi-supervised framework for crowd counting, which contains three major stages, i.e., labeling, training, and inference.

B. Problem Formulation

In active semi-supervised crowd counting, we are given a limited labeling budget (e.g., 10% of the training set). During labeling, we

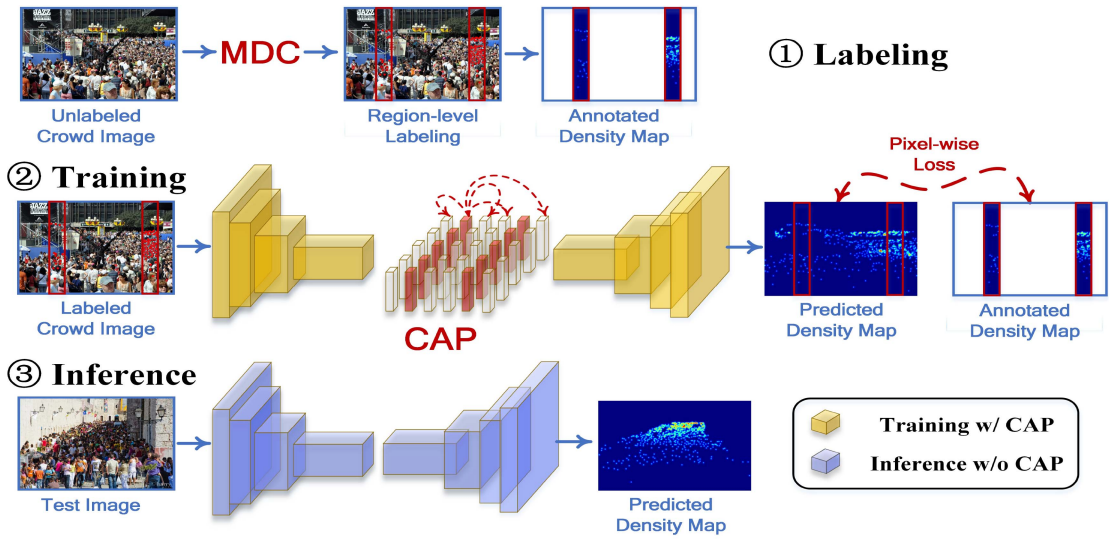

Fig. 2. Overview of the proposed semi-supervised crowd counting framework, which consists of three stages, i.e., labeling, training, and inference. At the labeling stage, we design a Multi-level Density-aware Cluster (MDC) strategy to annotate the representative regions only in each crowd image. After labeling, to leverage the rich unlabeled regions, a Crowd Affinity Propagation (CAP) module is introduced to supervise both the labeled and unlabeled regions via feature propagation by exploiting the deep feature affinities among individuals. Note that the CAP module can be removed at the inference stage without performance degradation, which makes it computationally free after training.

gradually gain access to a labeled set which is denoted as $\mathcal{S}_{\mathcal{L}} = \{(\mathbf{x}_i^l, \mathbf{y}_i^l)\}_{i=1}^{N_l}$, where $\mathbf{x}_i^l$ and $\mathbf{y}_i^l$ denote the i-th annotated crowd image and its corresponding label, i.e., a set of coordinates pointing out the positions of head centers. The labeling process involves initially and randomly annotated warm-up samples which are used to perform initial training of the crowd counter. Based on the initial crowd counter, we further select and annotate samples until we get the final labeled set $\mathcal{S}_{\mathcal{L}}$. Besides, the remaining unlabeled samples form an unlabeled set $\mathcal{S}_{\mathcal{U}} = \{(\mathbf{x}_j^u)\}_{j=1}^{N_u}$. Our goal is to utilize both sets to advance the crowd counting performance. Note that different from previous methods, we do not label all the regions in each image, so $\mathbf{x}_i^l$ and $\mathbf{x}_j^u$ are regions of crowd images in our context.

C. Crowd Counting Network

Crowd counting networks typically employ density maps as the intermediate output, which can be generated by convolving annotated head points with Gaussian kernels [4]

$$\mathcal{D}(\mathbf{z}) = \sum_{k=1}^{N} \delta(\mathbf{z} - \mathbf{z}_k) * G_{\sigma_k}(\mathbf{z}), \tag{1}$$

where $\mathbf{z}$ denotes each pixel in a crowd image $\mathbf{x}$. $\mathbf{z}_k$ represents the k-th annotated point (total N points). G_{σ_k} is a 2D Gaussian kernel with a bandwidth σ_k. Therefore, the crowd counting problem is converted to: $\mathcal{F} : \mathcal{I}(\mathbf{x}) \rightarrow \mathcal{D}(\mathbf{x})$, which learns a mapping from an image space $\mathcal{I}(\mathbf{x})$ to a density map space $\mathcal{D}(\mathbf{x})$. Following previous works [11], [12], [13], we employ a general and effective $\mathcal{F}$ based on CSRNet [24] to evaluate the effectiveness of proposed semi-supervised methods. To train $\mathcal{F}$, we adopt the pixel-wise euclidean loss to measure the distance between the annotated and estimated density maps

$$\mathcal{L}_{den}(\Theta) = \frac{1}{2\,M} \sum_{m=1}^{M} \|\mathcal{F}(\mathcal{I}_m; \Theta) - \mathcal{D}_m\|_2^2, \tag{2}$$

TABLE I
COMPARISON RESULTS OF DIFFERENT SPATIAL RATIOS OF THE LABELED REGIONS IN EACH CROWD IMAGE. THE EXPERIMENTS ARE CONDUCTED IN THE SHANGHAITECH PARTA DATASET WITH A 10% LABELING BUDGET. *:* DENOTES VERTICAL:HORIZONTAL. ∞:1 (OR 1:∞) REPRESENTS THE SPATIAL RATIO WITH THE HEIGHT (OR WIDTH) OF THE LABELED REGION EQUAL TO THAT OF THE ENTIRE IMAGE

Ratio	1:∞	1:4	1:2	1:1	2:1	4:1	∞:1 (ours)
MAE	95.9	95.2	93.8	93.3	92.4	91.3	**89.1**
RMSE	148.3	145.8	144.5	143.8	142.0	140.5	**137.5**

where Θ is the learnable parameters of $\mathcal{F}$. $\mathcal{I}_m$ is the m-th training image (total M images). $\mathcal{F}(\mathcal{I}_m; \Theta)$ and $\mathcal{D}_m$ denote the estimated and annotated density maps, respectively.

D. Representative Regions Selection Strategy

As we want to label more crowd images with diverse crowd scenes, we transfer the labeling budget to each crowd image. For example, if the budget is 10% of the entire dataset, we choose to label all the crowd images with 10% of each annotated. Then we come to the problem of how to find the representative regions in each crowd image.

Annotate More in the Vertical or Horizontal Direction? Regions in crowd images can be categorized into three types: dense, sparse, and background regions according to different crowd distributions. The representative regions in each crowd image should cover all the three types and have as large crowd density variations as possible given a limited labeling budget. As shown in Fig. 1, large crowd density variations usually appear in the vertical direction (e.g., from bottom to top) of each crowd image due to the surveillance camera perspective and imaging condition.

Therefore, when we are given a labeling budget in each crowd image, we should label a region which spreads more in the vertical direction than the horizontal direction (see Table I for an experimental comparison). Without loss of generality, we define the labeled region

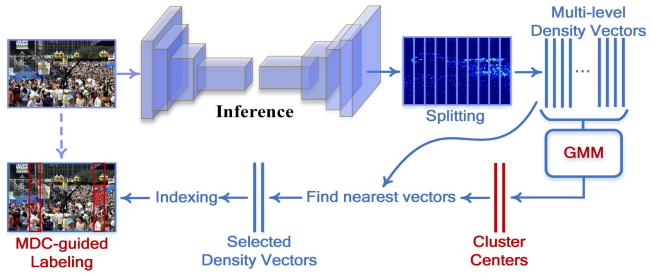

Fig. 3. Illustration of the proposed Multi-level Density-aware Cluster strategy for representative regions selection.

$\mathbf{x}_l \in \mathbb{R}^{H_l \times W_l}$ as a rectangular region of a crowd image $\mathbf{x} \in \mathbb{R}^{H \times W}$. As discussed above, H_l should be much larger than W_l (i.e., $H_l \gg W_l$). In the extreme case, $H_l = H$ and W_l varies according to the labeling budget. In practice, $\mathbf{x}_l$ may not be a continuous area in $\mathbf{x}$ and may contain n_l subregions. Therefore, the representative regions selection problem is simplified to determine the n_l subregions in $\mathbf{x}$.

Multi-level Density-aware Cluster (MDC) Strategy. The MDC strategy shown in Fig. 3 is designed to determine the n_l subregions in $\mathbf{x}$. First, we divide $\mathbf{x}$ into n_u rectangular subregions $\mathbf{x}_u = \{\mathbf{x}_u^1, \mathbf{x}_u^2, \ldots, \mathbf{x}_u^{n_u}\}$ with each $\mathbf{x}_u^j \in \mathbb{R}^{H \times W_u^j}$ and $W_u^j = Constant$. Then, our goal is to select n_l subregions $\mathbf{x}_l = \{\mathbf{x}_l^1, \mathbf{x}_l^2, \ldots, \mathbf{x}_l^{n_l}\}$ from $\mathbf{x}_u$. Different from the crowd density variations in the vertical direction, the crowd scene usually changes in the horizontal direction due to the large-view field of cameras. Therefore, we should select as many crowd scenes as possible in the horizontal direction. Based on the definition that the same crowd scene shares the same crowd density distributions along the horizontal direction, we propose a Multi-level Density-aware Cluster strategy to cluster the unlabeled regions $\mathbf{x}_u$ into multiple crowd scenes based on their multi-level density vectors. Specifically, to obtain density distributions of the unlabeled data, we first randomly label a few crowd images (e.g., 20% of the labeling budget) as warm-up samples to pretrain a crowd counter. Due to the existence of warm up samples, our method comes to one of *active* semi-supervised counting methods. For each unlabeled region $\mathbf{x}_u^j$ in a crowd image $\mathbf{x}$, we extract its predicted density maps $\mathbf{m}_u^j$ and calculate the multi-level density vector $\mathbf{v}_u^j$ of $\mathbf{m}_u^j$ as

$$\mathbf{v}_u^j = [\mathcal{V}^1(\mathbf{m}_u^j), \mathcal{V}^2(\mathbf{m}_u^j), \ldots, \mathcal{V}^L(\mathbf{m}_u^j)], \tag{3}$$

where $\mathcal{V}^L(\mathbf{m}_u^j)$ is the L-th level density vector defined as

$$\begin{aligned}\mathcal{V}^L(\mathbf{m}_u^j) = [\, &Sum(\mathbf{m}_u^j[H_L : H_L * 1, :]),\\ &Sum(\mathbf{m}_u^j[H_L * 1 : H_L * 2, :]), \ldots,\\ &Sum(\mathbf{m}_u^j[H_L * (L-1) : H_L * L, :])\,],\end{aligned} \tag{4}$$

where $Sum(\cdot)$ and $*$ denote the summation and multiplication operations. H_L is equal to H integrally divided by L. $\mathbf{m}_u^j[:, :]$ means a subregion of $\mathbf{m}_u^j$ where the former and latter dimensions are height and width. As the initial values in $\mathbf{v}_u^j$ have different scales, we normalize each of them to the same scale by $\mathcal{V}^k(\mathbf{m}_u^j)/L * k$ where k and $/$ denotes the k-th level and the division operation, respectively. With the multi-level design, $\mathbf{v}_u^j$ can express both the local and global crowd density distributions of $\mathbf{d}_u^j$.

Based on the calculated multi-level density vectors, we introduce a probabilistic cluster algorithm based on Gaussian Mixture Models (GMM) to cluster the unlabeled regions $\mathbf{x}_u$ into multiple crowd scenes. Details of the clustering algorithm are in the Supplementary Material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2022.3232712.

TABLE II
COMPARISON RESULTS OF DIFFERENT REPRESENTATIVE REGIONS SELECTION STRATEGIES IN THE SHANGHAITECH PARTA DATASET. EACH RESULT IS IN THE FORM OF MAE/RMSE. "MDC^K" AND "MDC^G" DENOTE MDC WITH K-MEANS AND GMM AS CLUSTERING METHODS, RESPECTIVELY. "EU" REPRESENTS ENSEMBLE-BASED UNCERTAINTY STRATEGY

Method	10%	20%	50%	90%
RANDOM	89.1/137.5	81.5/128.4	72.6/120.5	68.7/116.8
MAX	90.8/133.3	82.6/126.7	71.9/119.1	69.1/**116.3**
MDC^K	84.1/132.7	76.9/125.4	**71.0**/118.7	69.3/116.7
MDC^G	**83.3**/**132.1**	**76.4**/**125.2**	71.2/**118.4**	**68.5**/116.6
EU	86.4/133.9	78.1/126.9	72.2/119.5	68.9/117.2
EU + MDC^G	**82.8**/**131.4**	**76.6**/**124.7**	**70.6**/**117.5**	**68.7**/**116.4**

It is worth noting that Gaussian distributions predefined in GMM can facilitate the clustering process with an inductive prior. This implies that in natural crowd scenes, humans are typically distributed in a Gaussian or near-Gaussian manner. Multi-level density vectors are calculated based on density values in different regions of density maps, which express human distribution in each crowd scene. The Gaussian distribution is assumed on the multi-level density vectors, and thus on human distribution. Besides GMM, centroid-based clustering methods (e.g., K-means) are also good alternatives in our case considering the potential clustering methods should: i) specify the number of clusters in advance; ii) be effective and simple. We try both K-means and GMM in our experiments and find GMM can consistently perform better than K-means in the final counting performance (see Table II).

E. Crowd Affinity Propagation

To leverage the rich unlabeled regions, label propagation is a natural way. In crowd counting, pseudo labels mean pseudo density maps, which are typically error-prone as discussed in previous methods [11], [12], [13], [41]. Therefore, [11], [13], [41] utilize pseudo labels from *surrogate* tasks to supervise the feature extractor *only*. In this paper, instead of generating noisy pseudo labels, we propose a novel Crowd Affinity Propagation (CAP) module to directly supervise the unlabeled regions via feature propagation by exploiting deep feature affinities. The rationale behind this design is that prediction by comparison is more effective than direct prediction for the cases with only limited annotations. Crowd counting typically requires sufficient supervision to capture the diverse data distributions. However, the affinities between deep features in the latent space can infer whether they belong to the same class via relatively low-level semantics, e.g., similar color and texture.

Specifically, the CAP module contains two phases, e.g., forward propagation and backward propagation. In the forward propagation, deep features from the unlabeled regions are transferred to update those of the labeled regions by leveraging the feature affinities between them. Let $\mathbf{f}_u \in \mathbb{R}^{C \times H_u \times W_u}$ and $\mathbf{f}_l \in \mathbb{R}^{C \times H_l \times W_l}$ denote the features extracted by the crowd counter from the unlabeled and labeled regions in a crowd image. As $\mathbf{f}_u$ and $\mathbf{f}_l$ are extracted synchronously, they share the same number of channel dimensions C. The initial values in $\mathbf{f}_u$ and $\mathbf{f}_l$ may be very large or small, so we first normalize them as follows:

$$\mathbf{f}_u^n = \mathcal{S}(\mathbf{f}_u + \epsilon), \mathbf{f}_l^n = \mathcal{S}(\mathbf{f}_l + \epsilon), \tag{5}$$

where $\mathbf{f}_u^n$ and $\mathbf{f}_l^n$ are the normalized features. $\mathcal{S}(\cdot)$ denotes the softmax function along the channel dimension. ϵ is a small value to ensure

TABLE III
ABLATION STUDIES FOR THE CAP MODULE ON THE SHANGHAITECH PARTA DATASET. RESULTS ARE SHOWN IN THE FORM OF MAE/RMSE. "LP," "SST," AND "DST" REPRESENT DIRECT LABEL PROPAGATION [12], SEGMENTATION-BASED SURROGATE TASK [11], AND CLASSIFICATION-BASED SURROGATE TASK [13]

Method	10%	20%	50%
MDC	83.3/132.1	76.4/125.2	71.2/118.4
MDC + LP	84.7/135.3	76.9/125.1	71.5/118.3
MDC + SST	82.4/131.8	74.9/123.4	70.6/118.0
MDC + CST	82.8/131.5	74.2/122.9	70.5/117.5
MDC + CAP (train&infer)	**78.8**/127.9	**72.7**/**120.6**	**68.5**/116.1
MDC + CAP (train only)	79.6/**127.5**	73.2/121.3	69.2/**115.7**

stability. We reshape $\mathbf{f}_u^n$ to $\mathbb{R}^{C\times N_u}$ where $N_u = H_u \times W_u$ and $\mathbf{f}_l^n$ to $\mathbb{R}^{C\times N_l}$ where $N_l = H_l \times W_l$, and then $\mathbf{f}_u^n = \{\mathbf{f}_u^1, \mathbf{f}_u^2, \ldots, \mathbf{f}_u^{N_u}\}$ and $\mathbf{f}_l^n = \{\mathbf{f}_l^1, \mathbf{f}_l^2, \ldots, \mathbf{f}^{N_l}\}$ with each feature in $\mathbf{f}_u^n$ or $\mathbf{f}_l^n$ has C dimensions. Then we calculate the normalized similarity $\mathbf{s}_{ij}$ between each feature $\mathbf{f}_l^i$ of $\mathbf{f}_l^n$ and each feature $\mathbf{f}_u^j$ of $\mathbf{f}_u^n$ as follows:

$$\mathbf{s}_{ij} = \frac{exp(\mathbf{f}_l^i \cdot \mathbf{f}_u^j)}{\sum_{k=1}^{N_u} exp(\mathbf{f}_l^i \cdot \mathbf{f}_u^k))}. \tag{6}$$

The more similar $\mathbf{f}_l^i$ and $\mathbf{f}_u^j$ are, the higher $\mathbf{s}_{ij}$. After calculating the feature similarities, we propagate all the features of the unlabeled regions $\mathbf{f}_u^n$ to update each feature $\mathbf{f}_l^i$ of $\mathbf{f}_l^n$

$$\mathbf{f}_l^{i+} = \gamma \cdot \sum_{j=1}^{N_u} \mathbf{s}_{ij} \cdot \mathbf{f}_u^j + (1-\gamma) \cdot \mathbf{f}_l^i, \tag{7}$$

where $\mathbf{f}_l^{i+}$ is the updated i-th feature $\mathbf{f}_l^i$ of $\mathbf{f}_l^n$. γ is a learnable parameter to fuse the labeled and unlabeled features. After feature updating, the features of the labeled regions can also contain those of the unlabeled regions, which can be supervised by the labels of the labeled regions in the backward propagation. When the training procedure converges, we can optionally remove the CAP module from the crowd counter without performance degradation (see Table III for a detailed comparison) which means the proposed CAP module is computationally free at the inference stage.

F. Network Optimization

The proposed semi-supervised crowd counting framework contains three stages: (i) labeling, (ii) training with CAP, and (iii) inference without CAP. At the labeling stage, we first randomly label a small portion of the labeling budget as warm-up samples to pretrain a crowd counter for multi-level density vectors utilized in the MDC strategy. Then we label the remaining samples by the MDC strategy. After labeling, we train the crowd counter with the CAP module by all the labeled samples. At the inference stage, we remove the CAP module from the crowd counter and estimate density maps and crowd counts for any given crowd images.

IV. EXPERIMENTS

A. Implementation Details

As image resolutions in crowd counting datasets vary greatly, we set the batch size as 1 in all experiments. Without loss of generality, we empirically set the width W_u^j of each subregion $\mathbf{x}_u^j$ in Sec. 3.4 as 10% of the width of the corresponding crowd image and extract non-overlapping subregions in all datasets. As we want to label more crowd images with diverse individuals, we transfer the labeling budget to each crowd image. For example, if the budget is 10% of the entire dataset, we choose to label all the crowd images with 10% of each annotated. We utilize random cropping and horizontal flipping for data augmentation. Hyperparameters of the proposed method are chosen via cross validation. σ_k in (1) is a fixed bandwidth and is set as 4 for all datasets. L in (3) is set as 4 to balance the efficiency and computational cost. ϵ in (5) is set as 10^{-6} and γ in (7) is initialized as 0.2. Adam optimizer [42] is employed to optimize the crowd counting network with the initial learning rate as 10^{-8}. The experiments are conducted on a NVIDIA GTX 2080Ti GPU. More experimental details and results can be found in the supplementary materials, available online.

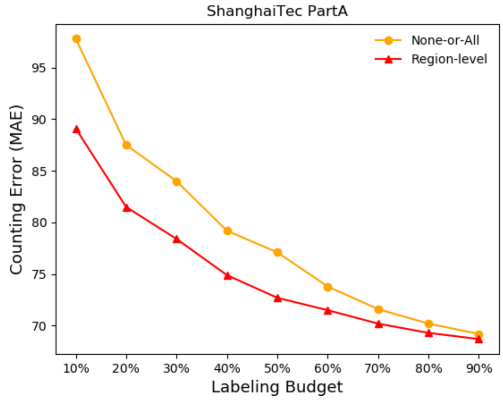

Fig. 4. Comparison results of the None-or-All and region-level labeling strategies with respect to different labeling budgets in the ShanghaiTech PartA dataset.

B. Ablation Studies

We conduct extensive ablation studies in the ShanghaiTech PartA dataset [4] to validate the effectiveness of the proposed semi-supervised crowd counting method.

Is region-level really better than None-or-All? To validate the effectiveness of the region-level labeling strategy, we randomly label images (for None-or-All strategy) and regions (for region-level strategy) and compare their performance with respect to different labeling budgets in Fig. 4. We can see that the region-level strategy can consistently outperform the None-or-All strategy. Note that the performance gain of the region-level strategy is magnified when the labeling budget gets small. This indicates that it is better to employ the region-level labeling strategy to annotate more crowd images with various crowd scenes, especially when the labeling budget is limited.

Annotate more in the vertical or horizontal direction? To validate the effectiveness of the vertical-first annotation strategy, we fix the labeling budget (i.e., 10%) in each crowd image and change the spatial ratios (i.e., vertical:horizontal) of the randomly labeled regions. The comparison results are shown in Table I. We can see that with more annotations in the vertical direction (e.g., the extreme case is ∞:1 where the height of the labeled region is equal to that of the entire image), the counting performance can be enhanced gradually, which confirms the effectiveness of the vertical-first annotation strategy considering the large crowd density variations caused by the camera perspectives.

Effectiveness of the MDC strategy for representative regions selection. Based on the verified region-level and vertical-first labeling strategy, we further evaluate the MDC strategy for representative regions selection. The experiments are conducted with respect to different region selection strategies, i.e., RANDOM, MAX, and MDC (i.e., MDC^K and MDC^G). MAX selects the regions with the maximum

numbers of people in each crowd image. for MDC, we fix the percentage of the randomly labeled samples (namely warm-up samples), i.e., 20% of the labeling budget.

The comparison results are shown in Table II. We can see that the MDC strategy outperforms the other two strategies consistently when the labeling budget is limited (e.g., 10% and 20%). When the labeling budget is abundant (e.g., 90%), the counting performance of the three labeling strategies is saturated without obvious differences. Note that the MDC strategy annotates less than the MAX strategy, e.g., when the labeling budget is 10%, MDC annotates 15,939 human heads while MAX annotates 20,401. This indicates that the proposed MDC strategy can achieve superior performance with less annotation burden. Comparing MDC^K with MDC^G, we can see MDC^G consistently outperforms MDC^K, which demonstrates the effectiveness of the inductive prior introduced by GMM for representative regions clustering.

In active learning [43], two aspects (i.e., uncertainty and diversity) are usually considered for annotation. In this paper, we focus on the diversity for annotating diverse crowd scenes, while general uncertainty-based methods usually introduce extra prediction heads to quantify uncertainty by predictions inconsistency [44], [45]. To compare with them, we add an extra prediction head and select the regions with the largest predictions inconsistency in each crowd image. The method is named as "EU" in Table II. We can see that our diversity-based strategy (i.e., MDC^G) can achieve better performance compared to the general uncertainty-based method (i.e., EU). Meanwhile, uncertainty is also feasible in our framework. Specifically, instead of selecting cluster centers as representative regions, we select the regions with the largest uncertainty in each cluster. The method is names as "EU + MDC^G" in Table II. We can see that our framework can integrate uncertainty and diversity which achieves better performance than each one of them.

Effectiveness of the CAP Module. Based on the well-performed MDC module, we evaluate the effectiveness of the proposed CAP module. Two variants are designed to explore the optimal setting of the CAP module, i.e., "CAP (train&infer)" and "CAP (train only)" in Table III. "CAP (train&infer)" means to add CAP both at the training and inference stages, while "CAP (train only)" removes CAP after the training stage. The comparison results are shown in Table III. We can see that with the CAP module, the counting performance of both "MDC + CAP (train&infer)" and "MDC + CAP (train only)" can be improved considerably, which demonstrates the effectiveness of exploited deep feature affinities to directly supervise the unlabeled regions. Besides, by the comparison between "MDC + CAP (train&infer)" and "MDC + CAP (train only)," we find that the CAP module can be removed at the inference stage without performance degradation, which indicates the proposed CAP module can enhance the counting performance efficiently without extra computational costs after training.

Besides, we compare the proposed CAP module with other strategies of exploiting unlabeled samples, i.e., direct Label Propagation (LP) [12], Segmentation-based Surrogate Task (SST) [11], and Classification-based Surrogate Task (CST) [13]. We can see in Table III that the proposed feature propagation strategy in CAP can achieve consistently superior performance compared to the other strategies.

Furthermore, we visualize the learned crowd affinities between labeled and unlabeled regions in Fig. 5. We can see the affinity maps can activate areas with the same semantics as the marked position, e.g, trees, skies, and humans with the same scale and illumination. This indicates that supervision signals can be effectively applied to the unlabeled regions via the explicit semantic exploration in CAP.

Whether the proposed MDC and CAP modules can benefit the existing None-or-All labeling strategy. To verify this, based on the None-or-All labeling strategy, we cluster images instead of regions for MDC and propagate features between images for CAP. The experiments

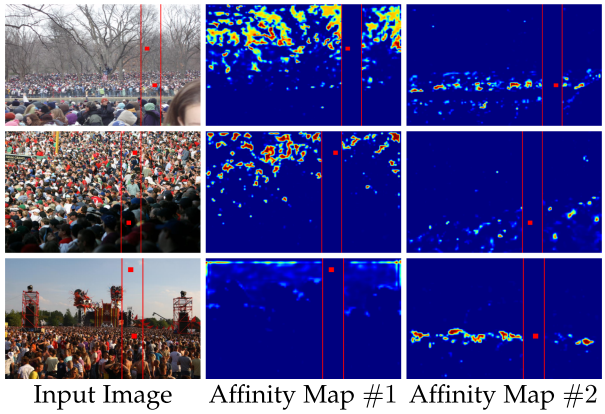

Fig. 5. Visualization of affinity maps in the CAP module. Labeled regions are enclosed between two red lines in each crowd image with two representative positions marked. Crowd affinities between each marked position and all the positions of unlabeled regions are illustrated in the latter two columns. Warmer colors mean higher values.

TABLE IV
FURTHER ABLATION STUDIES ON THE PROPOSED MAC AND CAP MODULES WHEN THEY ARE APPLIED TO THE EXISTING NONE-OR-ALL LABELING STRATEGY. THE EXPERIMENTS ARE CONDUCTED IN THE SHANGHAITECH PARTA AND UCF-QNRF DATASETS WITH 10% AND 20% LABELING BUDGETS, RESPECTIVELY

Method	STPart A		UCF-QNRF	
	MAE↓	RMSE↓	MAE↓	RMSE↓
None-or-All	98.4	160.1	151.3	259.0
None-or-All w/ MDC	90.5	144.2	139.4	242.7
None-or-All w/ CAP	91.3	147.8	141.0	245.4
None-or-All w/ MDC&CAP	**82.9**	**133.5**	**129.2**	**227.4**

are conducted in the ShanghaiTech PartA and UCF-QNRF datasets with 10% and 20% labeling budgets, respectively. We can see in Table IV that the proposed MDC and CAP modules can improve the performance of the existing None-or-All labeling setting, which implies a broader impact of the proposed semi-supervised modules.

Whether the proposed region-level labeling strategy can benefit existing semi-supervised counting methods. To verify this, we apply the proposed labeling strategy to existing semi-supervised counting methods, i.e., IRAST [11], AL-AC [13], and GP [12]. Specifically, for IRAST [11], we implement the surrogate segmentation task in the region level. For AL-AC [13], we actively select regions in each crowd image by PSSW and achieve region-based feature alignment. For GP [12], we implement the auxiliary Gaussian Process between labeled and unlabeled regions in each image. We conduct experiments in the ShanghaiTech PartA dataset with a 10% labeling budget. We can see in Table V that the proposed region-level labeling strategy can further boost the existing semi-supervised methods.

C. Comparison to State-of-the-Art Methods

In this section, we compare our method with state-of-the-art approaches, including MT [47], UDA [48], L2R [49], IRAST [11], AL-AC [13], GP [12], and SUA [41]. Among them, MT [47] and UDA [48] are the widely-used generic semi-supervised methods. L2R [49] is a self-supervised learning method which exploits unlabeled samples by ranking. IRAST [11], AL-AC [13], GP [12], and SUA [41] are semi-supervised crowd counting methods, which are based on the

TABLE V
FURTHER ABLATION STUDIES ON THE PROPOSED REGION-LEVEL LABELING STRATEGY WHEN IT IS APPLIED TO THE EXISTING SEMI-SUPERVISED METHODS. EXPERIMENTS ARE CONDUCTED IN THE SHANGHAITECH PARTA DATASET WITH A 10% LABELING BUDGET. AS THE METHODS DOES NOT RELEASE THEIR CODES, WE IMPLEMENT THEM AND SUMMARIZE THE PERFORMANCE IN THE "NONE-OR-ALL" COLUMN, WHICH ARE SLIGHTLY DIFFERENT FROM THE REPORTED ONES

Method	Type	None-or-All		Region-Level	
		MAE↓	RMSE↓	MAE↓	RMSE↓
IRAST [11]	S	87.3	145.4	**83.7**	**138.1**
GP [12]	S	95.2	152.7	**88.3**	**145.4**
AL-AC [13]	S^A	89.1	141.5	**84.0**	**133.9**

TABLE VI
COMPARISON WITH STATE-OF-THE-ART METHODS IN THE SHANGHAITECH PARTA [4] (DENOTED AS STPART A), SHANGHAITECH PARTB [4] (DENOTED AS STPART B), AND UCF-QNRF [46] DATASETS. THE LABELING BUDGETS ARE 10%, 10%, AND 20%, RESPECTIVELY. "S," "S^A," AND "F" DENOTE SEMI-SUPERVISED, ACTIVE SEMI-SUPERVISED AND FULLY-SUPERVISED METHODS, RESPECTIVELY. ITALIC NUMBERS REPRESENT RE-IMPLEMENTATION

Method	Type	STPart A		STPart B		UCF-QNRF	
		MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
CSRNet [24]	F	68.2	115.0	10.6	16.0	121.3	215.2
MT [47]	S	94.5	156.1	15.6	24.5	145.5	250.3
UDA [48]	S	93.8	157.2	15.7	24.1	144.7	255.9
L2R [49]	S	90.3	153.5	15.6	24.4	148.9	249.8
IRAST [11]	S	86.9	148.9	14.7	22.9	135.6	233.4
AL-AC [13]	S^A	87.9	139.5	**12.7**	20.4	*131.4*	*229.7*
SUA [41]	S	85.1	–	–	–	–	–
Ours	S^A	**79.6**	**127.5**	**12.7**	**20.3**	**128.6**	**226.4**

None-or-All labeling strategy. All the comparison methods are based on CSRNet [24] with a VGG16 backbone network. The comparison results are shown in Table VI.

We can see in Table VI that the semi-supervised counting methods (i.e., IRAST [11], AL-AC [13], and SUA [41]) can achieve superior performance compared to the general semi-supervised methods (i.e., MT [47] and UDA [48]). However, they are still far from the fully-supervised CSRNet model. Differently, our method can effectively narrow down the performance gap and enhance the state-of-the-art semi-supervised counting performance by a large margin. AL-AC [13] seems comparable to the proposed method in ShanghaiTech PartB dataset. To show more comparisons with AL-AC, we implement it and further report their performance on UCF-QNRF. We can see that our method can outperform AL-AC by a large margin in the ShanghaiTech PartA and UCF-QNRF datasets. Following GP [12], we also annotate 5% of the entire training set in the ShanghaiTech PartA and UCF-QNRF datasets. The performance of our method is 89.7/135.6 (PartA) and 138.9/247.1 (UCF-QNRF) in terms of MAE/RMSE which are much better than 111/159 and 171/293 of GP (the numbers are from Table VI of GP where CSRNet is utilized as the backbone network).

More Recent Datasets. To verify our method in more recent datasets (i.e., JHU-CROWD++ [50] and NWPU [51]), we report the performance in Table VII. As only SUA [41] explores the two datasets, we fix the labeling budget (i.e., 50%) following SUA and compare with it in Table VII. We can see that our method can also achieve superior performance in these two larger and more challenging datasets.

TABLE VII
COMPARISON RESULTS IN JHU-CROWD++ [50] AND NWPU [51] DATASETS. AS ONLY SUA [41] EXPLORES THE TWO DATASETS, WE FOLLOW IT TO SET THE LABELING BUDGETS AS 50%

Method	Type	JHU-CROWD++		NWPU	
		MAE↓	RMSE↓	MAE↓	RMSE↓
SUA [41]	S	80.7	290.8	111.7	443.2
Ours	S	**80.2**	**287.5**	**109.3**	**438.1**

V. CONCLUSIONS AND LIMITATIONS

In this work, we propose to break the labeling chain of previous methods and make the first attempt to reduce spatial labeling redundancy for effective semi-supervised crowd counting. Specifically, we analyze the region representativeness from both the vertical and horizontal directions, and formulate the representative regions as cluster centers of Gaussian Mixture Models based on their multi-level density vectors. Additionally, we design a Crowd Affinity Propagation (CAP) module to directly supervise the unlabeled regions via feature propagation without the error-prone pseudo label generation. Extensive experiments on widely-used benchmarks demonstrate that our method outperforms previous best approaches by a large margin.

Our main idea is to annotate the representative *regions only* in each crowd image. One limitation is that we still need to annotate all the human heads in each representative region. Can we annotate representative *humans only* in each crowd image without choosing regions? The setting is more challenging as in this way we do not have fully-annotated regions to generate density maps for supervision. As we are the first to break the None-or-All labeling paradigm, we leave this as a potential future work.

REFERENCES

[1] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 367–386, Mar. 2015.

[2] C. C. Loy, K. Chen, S. Gong, and T. Xiang, "Crowd counting and profiling: Methodology and evaluation," in *Modeling, Simulation and Visual analysis of Crowds*, Berlin, Germany: Springer, 2013, pp. 347–382.

[3] G. Gao, J. Gao, Q. Liu, Q. Wang, and Y. Wang, "CNN-based density estimation and crowd counting: A survey," 2020, *arXiv:2003.12783*.

[4] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 589–597.

[5] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5094–5103.

[6] S. Bai, Z. He, Y. Qiao, H. Hu, W. Wu, and J. Yan, "Adaptive dilated network with self-correction supervision for counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4593–4602.

[7] J. Wan and A. Chan, "Modeling noisy annotations for crowd counting," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 3386–3396.

[8] M. Shi, Z. Yang, C. Xu, and Q. Chen, "Revisiting perspective information for efficient crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7271–7280.

[9] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Leveraging unlabeled data for crowd counting by learning to rank," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7661–7669.

[10] D. B. Sam, N. N. Sajjan, H. Maurya, and R. V. Babu, "Almost unsupervised learning for dense crowd counting," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8868–8875.

[11] Y. Liu, L. Liu, P. Wang, P. Zhang, and Y. Lei, "Semi-supervised crowd counting via self-training on surrogate tasks," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 242–259.

[12] V. A. Sindagi, R. Yasarla, D. S. Babu, R. V. Babu, and V. M. Patel, "Learning to count in the crowd from limited labeled data," 2020, *arXiv:2007.03195*.

[13] Z. Zhao, M. Shi, X. Zhao, and L. Li, "Active crowd counting with limited supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 565–581.

[14] B. Tan, J. Zhang, and L. Wang, "Semi-supervised elastic net for pedestrian counting," *Pattern Recognit.*, vol. 44, no. 10/11, pp. 2297–2304, 2011.

[15] C. C. Loy, S. Gong, and T. Xiang, "From semi-supervised to transfer counting of crowds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2256–2263.

[16] Q. Zhou, J. Zhang, L. Che, H. Shan, and J. Z. Wang, "Crowd counting with limited labeling through submodular frame selection," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1728–1738, May 2019.

[17] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–11.

[18] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2547–2554.

[19] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 1324–1332.

[20] Z. Lin and L. S. Davis, "Shape-based human detection and segmentation via hierarchical part-template matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 604–618, Apr. 2010.

[21] M. Wang and X. Wang, "Automatic adaptation of a generic pedestrian detector to a specific traffic scene," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3401–3408.

[22] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 247–266, 2007.

[23] Y. Liu et al., "Crowd counting via cross-stage refinement networks," *IEEE Trans. Image Process.*, vol. 29, pp. 6800–6812, 2020.

[24] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1091–1100.

[25] L. Chai, Y. Liu, W. Liu, G. Han, and S. He, "Crowdgan: Identity-free interactive crowd video generation and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2856–2871, Jun. 2022.

[26] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4031–4039.

[27] X. Jiang et al., "Crowd counting and density estimation by trellis encoder-decoder networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6126–6135.

[28] V. A. Sindagi and V. M. Patel, "HA-CCN: Hierarchical attention-based crowd counting network," *IEEE Trans. Image Process.*, vol. 29, pp. 323–335, 2019.

[29] X. Jiang et al., "Attention scaling for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4705–4714.

[30] Y. Miao, Z. Lin, G. Ding, and J. Han, "Shallow feature based dense attention network for crowd counting," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11765–11772.

[31] A. Zhang et al., "Relational attention network for crowd counting," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6787–6796.

[32] A. Zhang et al., "Attentional neural fields for crowd counting," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5713–5722.

[33] J. Wan and A. Chan, "Adaptive density map generation for crowd counting," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1130–1139.

[34] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6141–6150.

[35] L. Liu, H. Lu, H. Zou, H. Xiong, Z. Cao, and C. Shen, "Weighing counts: Sequential crowd counting by reinforcement learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 164–181.

[36] V. Ranjan, B. Wang, M. Shah, and M. Hoai, "Uncertainty estimation and sample selection for crowd counting," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 375–391.

[37] M.-H. Oh, P. A. Olsen, and K. N. Ramamurthy, "Crowd counting with decomposed uncertainty," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11799–11806.

[38] Z. Yan et al., "Perspective-guided convolution networks for crowd counting," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 952–961.

[39] Y. Yang, G. Li, Z. Wu, L. Su, Q. Huang, and N. Sebe, "Weakly-supervised crowd counting learns from sorting rather than locations," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–17.

[40] D. Liang, X. Chen, W. Xu, Y. Zhou, and X. Bai, "Transcrowd: Weakly-supervised crowd counting with transformer," 2021, *arXiv:2104.09116*.

[41] Y. Meng et al., "Spatial uncertainty-aware semi-supervised crowd counting," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 15549–15559.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[43] P. Ren et al., "A survey of deep active learning," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–40, 2021.

[44] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, "The power of ensembles for active learning in image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9368–9377.

[45] S. Lee, M. Amgad, M. Masoud, R. Subramanian, D. Gutman, and L. Cooper, "An ensemble-based active learning for breast cancer classification," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2019, pp. 2549–2553.

[46] H. Idrees et al., "Composition loss for counting, density map estimation and localization in dense crowds," in *Proc. Eur. Conf. Comput. Vis.*, 2018.

[47] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 544–559.

[48] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6256–6268.

[49] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Exploiting unlabeled data in CNNs by self-supervised learning to rank," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1862–1878, Aug. 2019.

[50] V. A. Sindagi, R. Yasarla, and V. M. Patel, "Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1221–1231.

[51] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-Crowd: A large-scale benchmark for crowd counting and localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2141–2149, Jun. 2021.