Glance to Count: Learning to Rank with Anchors for Weakly-supervised Crowd Counting

Zheng Xiong[1,2], Liangyu Chai[1,2], Wenxi Liu[3], Yongtuo Liu[4], Sucheng Ren[2] and Shengfeng He[*2]

[1]South China University of Technology, [2]Singapore Management University
[3]Fuzhou University, [4]University of Amsterdam

Abstract

Crowd image is arguably one of the most laborious data to annotate. In this paper, we aim to reduce the massive demand for densely labeled crowd data, and propose a novel weakly-supervised setting, in which we leverage the binary ranking of two images with high-contrast crowd counts as training guidance. To enable training under this new setting, we convert the crowd count regression problem to a ranking potential prediction problem. In particular, we tailor a Siamese Ranking Network that predicts the potential scores of two images indicating the ordering of the counts. Hence, the ultimate goal is to assign appropriate potentials for all the crowd images to ensure their orderings obey the ranking labels. On the other hand, potentials reveal the relative crowd sizes but cannot yield an exact crowd count. We resolve this problem by introducing "anchors" during the inference stage. Concretely, anchors are a few images with count labels used for referencing the corresponding counts from potential scores by a simple linear mapping function. We conduct extensive experiments to study various combinations of supervision, and we show that our method outperforms existing weakly-supervised methods by a large margin without additional labeling effort. The code is available at `https://github.com/pandaszzzzz/CCRanking`.

1. Introduction

Crowd counting aims to automatically count the number of individuals in images and has been widely applied in many areas, *e.g.*, video surveillance, traffic estimation, and congestion control. Most recent approaches [58], [59], [5], [19] rely mainly on fully-supervised annotation for individuals in the crowd (*i.e.*, placing a dot at the center of each individual) to estimate crowd density. Yet, such an annotation process is extremely time-consuming and laborious. Especially for extremely dense scenarios, it is almost senseless to manually label over-heaped dots just for the purpose of representing crowd density in a scene. Such a tedious annotation process hinders the scale and diversity of crowd datasets and thus slows down the development of this area.

*Corresponding author. Email: shengfenghe@smu.edu.sg.

$\mathcal{A}$ $\mathcal{B}$ $\mathcal{C}$

Figure 1: For the crowd images with more than twice the number of differences, humans can readily tell which one is of more people. We can assert that the number of people is: $count(\mathcal{A}) < count(\mathcal{B}) < count(\mathcal{C})$ intuitively. (The exact numbers: $count(\mathcal{A}) = 254$, $count(\mathcal{B}) = 580$, $count(\mathcal{C}) = 1202$. Please zoom in for a better view.) We aim to resolve the crowd counting problem by solely relying on ranking two images with high contrast in crowd counts.

Recent work [56] revisits the regression-based counting method that ignores the exact individual locations and directly maps a crowd image to its crowd counts. However, the problem of annotation remains unsolved, as ground-truth crowd counts are required for training, which thus cannot prevent annotators from strenuously pinpointing each individual in the images. Besides, some approaches [10], [2] aim to bypass dense annotations with alternative interactions. Considering each annotated object in an image is atomic and equivalent, they require a few individual annotations instead of accurate locations of all objects. These methods are promising to relieve the annotation efforts while still achieving good performance, but they do not radically solve the problem.

To address the above concern, we rethink the way of governing crowd counting models. Intuitively, directly estimating the crowd counts in an image is a challenging task even for a human expert. But it is much easier to sense the relative density for a few crowd images with great contrast in population sizes. For example, for the three crowd images in Fig. 1, we can easily tell which one has the largest scale crowd. Thus, this observation sheds light on a novel methodology of supervising the crowd counting models us-

ing *ranking labels* for any pair of crowd images, where each ranking label indicates which one of the image pair contains more persons. Compared to the existing annotation process, annotating ranking labels for images with a large contrast in crowd sizes is an almost effortless task for human annotators, so establishing larger diverse crowd datasets becomes possible, which will be of great value to the community. To this end, it boils down to the problem of how to effectively leverage the ranking labeled image pairs to supervise crowd counting models in such a weakly-supervised manner.

To exploit the ranking labels of crowd images, we propose a novel Siamese Ranking Network (SRN). Specifically, instead of directly estimating the crowd size of each image, a pair of crowd images are fed into two separate yet weight-sharing deep networks to predict their potentials that indicate the ordering of the counts in crowd images. For instance, in Fig. 1, the potential of $\mathcal{C}$ should be higher than those of $\mathcal{A}$ and $\mathcal{B}$. During the training phase, the magnitude relationship between a pair of predicted potentials can be supervised by the corresponding ranking labels of the crowd image pair. To this end, the purpose of model optimization is to assign appropriate potentials for all crowd images, which makes their orderings of potentials consistent with the ranking labels. Thus, the predicted potentials will be innately and positively correlated with the number of objects in crowd images, so they can be further applied to regress the absolute crowd counts. We construct a large number of ranking labels corresponding to pairs of crowd images as the training set for training our proposed Siamese network.

However, the learned Siamese network can only produce potential scores as no crowd count labels are involved in training. To combat this problem, in the inference stage, we leverage a few images with crowd count labels as the reference, denoted as *counting anchor set*, to disentangle the relationship between potential scores and actual crowd counts. Thanks to the positive correlation between the potentials and the ground-truth crowd counts, we establish a linear mapping function to fit the potential and counts of anchors, and we apply it to estimate the crowd count of the query image. Our proposed annotation scheme is far more facile than the standard point-based annotation, especially in challenging crowd scenes, but we achieve comparable performances to those trained with stronger supervision. Extensive experiments study various combinations of supervision in training, and we show that our model is superior to state-of-the-art weakly-supervised methods, even better than those trained with point-based labels.

Our contributions are three-fold:

- We introduce a novel weakly-supervised crowd counting setting, which can reduce labeling costs largely and be notably beneficial to dense and congested crowds.
- We propose a simple but effective Siamese-training method and utilize the anchoring mechanism to estimate crowd counts and we verify its effectiveness in this task.
- Extensive experiments conducted on several challenging benchmarks study various combinations of supervision. We demonstrate that our method outperforms state-of-the-art weakly-supervised methods without any extra labeling effort.

2. Related Work

Fully-supervised Crowd Counting. In recent years, deep learning based methods [48], [59], [53], [25], [19] have attracted much attention in computer vision for crowd counting. The crowd counting methods mainly include detection-based methods and regression-based methods. For the detection-based methods, Stewart *et al*. [45] propose to learn person the detector relying on bounding box annotations to count. [30] only requires point supervision to detect the human heads and count them in crowds simultaneously. However, it is difficult to accurately detect heads or bodies in extremely dense and congested crowd scenes, and that always degrades counting performance.

Therefore, the mainstream idea is to train deep CNN networks for density regression. CNN-based regression methods learn a mapping from semantic features to a density map and predict the total count. The main issue of regression-based counting tasks is the huge variation of instance scales. To tackle scale variations, employing multiple receptive fields is effective to learn from people of various sizes. For instance, several works [58], [59], [38], [43], [8], [36] employ multi-column networks to obtain local or global contextual features to handle scale variations. [5], [26], [15] utilize inception blocks to acquire different receptive fields. Several approaches [19], [6] combine the semantic features with dilated convolution for density estimation. Meanwhile, some works [34], [60], [57], [49] introduce the attention mechanism which is effective in extracting foreground features. Considering performance gain from extra supervision, perspective maps [27], [41], [55] and depth maps [20] are delivered to bring more scale guidance. On the other hand, combining with high-level tasks, *i.e.*, localization [31], [24], segmentation [42], depth prediction [60], can provide more accurate location labels for density regression and boost the counting with extra semantic information.

However, all the above CNN-based methods require a large number of labels during training, and annotating the crowd counting dataset is a labor-intensive and time-consuming task.

Weakly-/Un-/Self-Supervised Crowd Counting. There are some weakly-, un-, or self-supervised counting methods proposed with the consideration of relieving the labeling burden. In the weakly-supervised setting, most

methods are regression-based and adopt the image-level count label as the weak supervision signal for training. Idrees *et al*. [13] leverage Fourier Analysis as feature extraction mechanisms to predict total counts. The work in [47] applies the Gaussian process as a weakly-supervised solution for crowd counting. For the CNN-based methods, [56] proposes a soft-label sorting network to strengthen the supervision of crowd numbers beyond the original counting network, and [17] is a semi-supervised method combining a few location-level labels with count-level annotations. [32] focuses on highly confident regions while addressing the noisy supervision from unlabeled data as well in a semi-supervised manner. Although the above count-level methods are "weakly-supervised", the time spent for annotation actually is not remarkably reduced, while just the number of labels is reduced. Besides, Sam *et al*. [37] develop an autoencoder to achieve crowd counting in an almost unsupervised manner, and only a few parameters are updated when training. By matching statistics of the distribution of labels, they propose a completely self-supervised training paradigm without using any annotated image in [35]. But the performance of unsupervised methods still exists a large gap with fully-supervised works. Besides, it is known that deep CNN-based crowd counting methods usually struggle with the overfitting problem due to existing small datasets and their limited variety. To ease the overfitting problem, Wang *et al*. [51] explore generating synthetic crowd images to reduce the burden of annotation and alleviate overfitting.

With the rapid development of Transformers [46] in the field of computer vision, various methods such as TransCrowd [21], CrowdMLP [50], CrowdFormer [39], and HACC [18] have employed the Transformer architecture. Among them, TransCrowd [21] utilizes a Transformer encoder based on ViT [9] and employs a regression head for weakly supervised crowd counting. CrowdFormer [39] builds upon this by incorporating multi-scale fused features. However, these methods still require massive count labels and are not effective in the weakly-supervised setting.

Compared with the prior weakly-supervised methods, our proposed training settings based on ranking labels can effectively reduce the labeling burden, while maintaining the state-of-the-art performance.

Learning to rank. Different from the standard machine learning tasks of regression or classification, the ranking tasks are not with precise ground-truth metric targets or class labels as supervision. These works are designed to handle with ordered ranks, *i.e.*, which may be from human preferences, to predict the ordinal rank or relevant metric. There are many learning-to-rank works proposed in the literature. Some of the approaches, such as Ranking SVM [12], RankBoost [11], and RankNet [4], focus on a pair of instances to learn their ranking function by minimizing the loss functions. [40] applies learning-to-rank to large-scale datasets with the Stochastic Gradient Descent method.

In contrast to the above works for predicting ranks, additional numerical references or other auxiliary supervision is required if we will make a prediction on the number of count. Liu *et al*.[28] propose a learning-to-rank self-supervised strategy for utilizing available unlabeled images. And they show the ranking can be used as a proxy task for some regression tasks to solve the problem of limited size in existing datasets. However, relying the ranking of cropped images, which are cropped from the same source image, suffers the influence of similar paired patterns.

In this paper, the proposed ranking strategy works on pairwise instances of different images, similar to ranking SVM [12]. The setting can be free from the aforementioned problems from the same source pattern. And our method alleviate both overfitting problem and intensive annotation burden. The performance in a weakly-supervised setting is comparable to location-level supervised methods, such that can be feasibly applied in practical applications.

3. Proposed Method

3.1. Problem Formulation

Given a set of crowd images $X = \{x_1, x_2, ..., x_n\}$, where x_i is the sample of crowd images, and $x_i \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the height, width, and number of channels respectively, our goal is to predict the object count y_i of a crowd image x_i. The existing works achieve the goal by learning a mapping function $M : X \to Y$ to predict the labels, *i.e.*, $Y = \{y_1, y_2, ..., y_n\}$. Different from the previous practice, we utilize *ranking labels* for supervision instead of count labels. Count labels refer to the object number of crowd images, while ranking labels are binary (i.e. $\{-1, 1\}$), each of which indicates the size relationship between the counts of two crowd images.

Formally, the crowd dataset X can be expanded as a set P of ranking labeled pairs. Each element of the ranking pair set P in the crowd dataset X is denoted as a tuple $\langle x_i, x_j, q_{i,j}\rangle, 1 \le i < j \le n$, where the binary ranking label $q \in \{-1, 1\}$. When the crowd counts of x_i and x_j are subject to $y_i \gg y_j$, their size relationship can be easily identified, which means that the crowd image x_i ranks higher than x_j, so that the corresponding label $q_{i,j}$ is set to 1 and vice versa.

In favor of learning the ranking relationship among a set of crowd images, we introduce the *crowd counting potential* v that is positively correlated with the actual crowd count y. Thus, the goal is to learn an estimating function $f(x; w)$ to predict the potentials v for all the crowd images with ranking labels, which aims to ensure the ordering of the assigned potentials obeys the ranking labels. After that, the estimation function can be further used to predict the potential for

any query image and the anchoring mechanism can map its potential to the absolute crowd count.

3.2. Labeling Ranking Pairs

For a pair of unlabeled crowd images $\{x_1, x_2\}$, if it is observed that the crowd number y_1 is definitely much greater than y_2, the ranking label $q_{i,j}$ is set to 1, and vice versa. If the relationship is hard to identify, no label will be added to the ranking pair set P. Regarding various crowd scenarios, there are different strategies to label ranking pairs.

First, we can exploit the images from public benchmarks to train our model. For most existing crowd counting benchmarks (*e.g.*, ShanghaiTech [59], UCF-QNRF [14]), they can provide the dense annotations of object locations for crowd images and the count labels can thus be acquired without effort, which means we can easily convert them into ranking labels. Generally, for crowd dataset X, the size of all ranking pairs $|P_{all}|$ is $O(n^2)$ (where n is the size of the dataset), in which any two images could be formed as a ranking pair. However, the number of ranking labels is much less than n^2, since we are only allowed to manually label the obviously distinguishable pairs of crowd images. In addition, considering that the ranking relationship is transitive, if the ranking labels, $\langle x_A, x_B, 1\rangle$ (i.e. $y_A > y_B$) and $\langle x_B, x_C, 1\rangle$ (i.e. $y_B > y_C$), already exist in the annotated set, then $\langle x_A, x_C, 1\rangle$ or $\langle x_C, x_A, -1\rangle$ will be automatically added to the annotated set as shown in Fig. 2. Particularly, automatic labeling can be realized by detecting the connectivity in a directed acyclic graph, $G = (V, E)$. In specific, V denotes the set of vertices, which includes all the images that occurred in the ranking pairs. E denotes the set of arcs, where $\langle x_i, x_j, q_{i,j}\rangle, q_{i,j} = 1$ refers to an arc $i \rightarrow j$ from the vertex i to th vertex j on the graph G .

3.3. Pairwise Ranking Model

Our proposed Siamese ranking network architecture is illustrated in Fig. 2. Specifically, we employ a deep Siamese Network [7] as the ranking model for pairwise crowd images, which consists of two branches of networks that share weights. The input of our model is a ranking pair $\{x_1, x_2\}$ from the set P. Each branch of the network is fed with one of the image pairs and outputs the corresponding potentials v_1 and v_2. We leverage a Transformer-based feature extractor architecture, PVTv2 [52], as the backbone of the Siamese network. The backbone is composed of four stages to extract feature maps of different scales. The feature map obtained from the i-th stage exhibits dimensions that are reduced by a scale factor of 2^{i+1} compared to the original input image size. After obtaining feature maps at four different scales, we propose a Multi-Scale Feature Fusion module to generate a feature map that incorporates multi-scale information. Due to their varying scales and channel numbers, we initially standardize their scales using a Global Average Pooling and subsequently ensure uniform channel numbers through a fully connected layer. Finally, we can merge them together through summation, resulting in the final extracted feature. We design a Potential Decoder to map the feature maps extracted by Transformer Backbone to the potential values of the scene. Here, we use three fully connected layers to reduce the dimension of the features to 1. To supervise the ranking labels, we propose to associate the predicted potentials of two network branches with the corresponding ranking labels. Inspired by Ranking Support Vector Machine [12], we can train the network by minimizing the ranking hinge loss as follows:

$$\mathcal{L}_{rank} = \sum_{x_i, x_j} max(0, f(x_j; w) - f(x_i; w) - \mathcal{M}),$$
$$\text{s.t. } \langle x_i, x_j, 1\rangle \text{ or } \langle x_j, x_i, -1\rangle \in P, \tag{1}$$

where $\mathcal{M}$ is a hyper-parameter that indicates a margin to maintain the potential difference between paired images and w refers to the network weights.

By analyzing the population distribution within the dataset, it has been observed that in many scenes, the crowd tends to be denser in the lower half of the image while sparser in the upper half. This phenomenon can be attributed to the influence of gravity, as it is reasonable for people to stand on the ground rather than in the air. However, this tendency can lead to model errors when the model excessively relies on this characteristic of population distribution. In order to enhance the robustness of the model, the original input $\{x_1, x_2\}$ is modified to $\{x_1, \hat{x}_2\}$, which means that there is a 0.5 chance for image x_2 to upside-down vertical flipping. Subsequently, an Upside-Down MLP is incorporated after the backbone network to detect whether the x_2 features have flipped. Cross-entropy is used to supervise the flip label and the prediction of MLP.

$$\mathcal{L}_{cls} = \sum_{i} \langle y_i, \log(g(x_i; w))\rangle. \tag{2}$$

By introducing this auxiliary classification task, the encoder of our model is able to extract more accurate features across various scenarios.

Specifically, during training, we want the model to discard the redundant samples that are easy to discriminate and cannot contribute to the model optimization. To do so, we set a *hard sample filter* which requires the potentials v_i and v_j, inferred from $\{x_i, x_j\}$, should be subject to the condition $\frac{v_i}{v_j} < \xi$ (or $\frac{v_j}{v_i} < \xi$), where ξ is a predefined threshold to determine if the ranking labels should be abandoned.

3.4. Anchoring Mechanism

During inference, for any query crowd image $\hat{x}$, its predicted potential can be used to regress its crowd count. Thus, it is intuitive to compare the potential $\hat{v}$ of the query image (i.e., $\hat{v} = f(\hat{x})$) against a set of exemplar crowd images with their ground-truth counts known.

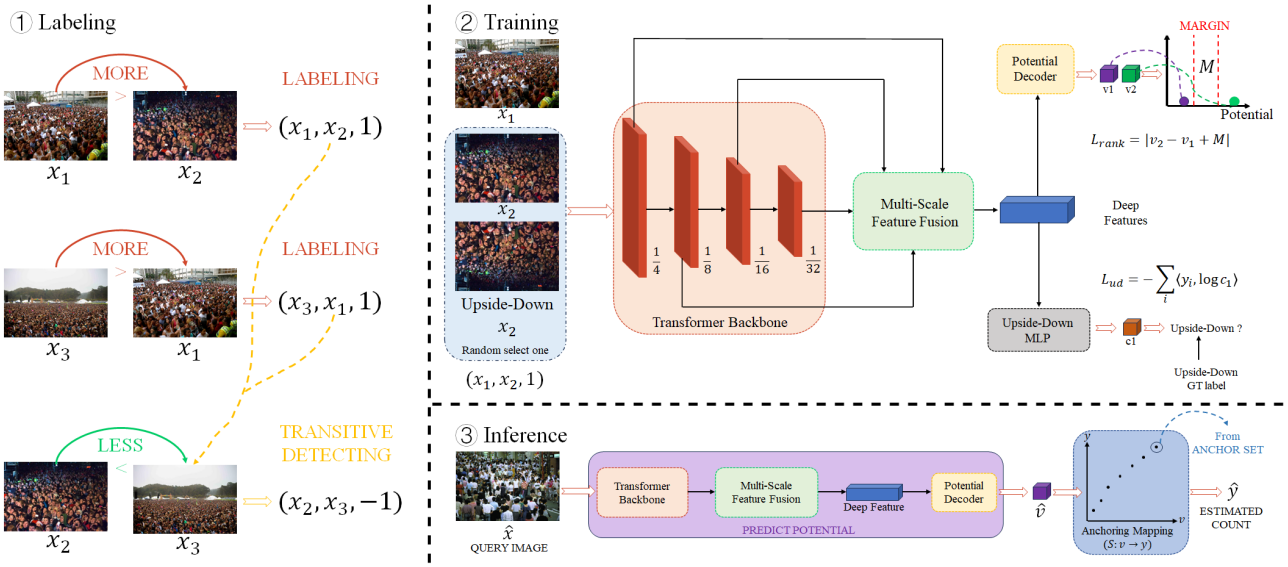

Figure 2: Our framework consists of three stages. (1) Constructing a DAG to detect transitive relations amongst crowd images, i.e., automatic labeling, which establishes a crowd dataset with ranking labels. (2) Training the Siamese Ranking Network on ranking labels to minimize the ranking hinge loss and predict potentials (e.g. v_1, v_2) for crowd images, where the parameter $\mathcal{M}$ denotes the margin of the loss. (3) On the inference stage, the count of a crowd image $\hat{x}$ can be estimated by mapping the predicted potential $\hat{v}$ to actual crowd count $\hat{y}$.

In particular, we introduce a counting anchoring mechanism. We denote the labeled exemplar images as *counting anchor set*. The anchor set includes a set of crowd images sampled from the training set, and their counts are specifically annotated and distributed over a large counting range, *e.g.*, 30 $\sim$ 3,000. Thus, we can estimate the scaling mapping function $S : v \rightarrow y$ that projects the predicted potentials of the images in the anchor set to crowd counts via linear regression, as shown in Fig. 2 ③. We use a L1 loss to fit the mapping between potential values and counts:

$$\mathcal{L}_{reg} = \frac{1}{n} \sum_{i}^{n} |y_i - v_i|. \tag{3}$$

By learning to rank, our model effectively captures crowd characteristics within the scene. The linear function S then transforms this latent information into precise counting, yielding heightened precision. Our experiments in Sec. 4 demonstrate a distinct positive relationship between the model's predicted potential values and corresponding counting labels. This affirms the viability of utilizing a linear function for fitting.

3.5. Ranking v.s. Regression

The reasons for choosing a ranking-based scheme over regression-based methods are twofold. First, compared to the point-based annotation, ranking labels are much easier to obtain, i.e., annotators only need to label "close", "less than", and "greater than" for any two crowd images. Second, the ranking problem is simpler to solve than regression-based methods, because ranking relationships are invariant to any image-level geometrical transformation, which can thus improve the model's robustness. On the contrary, optimizing the regression model is difficult and tends to get stuck with local minima. Although it is easy to obtain the optimal solution for the ranking-based formulation, it may not always be optimal for crowd counting. This is because the theoretical upper bound of the ranking optimization objective is easy to approach. It seems that simply learning a good ranking model is insufficient for counting tasks, and a suboptimal regression model can also give a perfect ranking. But a model with perfect ranking performance may yield poor regression performance.

Therefore, it is a trade-off for choosing ranking or regression. To further improve our model, we propose to integrate regression and ranking-based schemes. To this end, crowd counting can be achieved in such a hybrid weakly-supervised setting. The optimization objective consists of three terms, *i.e.*, pairwise ranking loss $L_{rank}(P; w)$, regression loss $L_{reg}(D; w)$, and classification loss $L_{cls}(T; w)$. In each iteration during training, there will be a crowd image randomly selected from D and a ranking pair randomly selected from P as network input. Formally, the hybrid optimization loss is defined as below:

$$\mathcal{L}^{++} = \min_{w} \mathcal{L}_{rank}(P; w) + \alpha_1 \mathcal{L}_{reg}(D; w) + \alpha_2 \mathcal{L}_{cls}(T; w), \tag{4}$$

where P denotes the dataset encompassing pairs of images with ranking labels, D signifies the anchor set consisting of

crowd images with counting labels employed exclusively for regression, and T encompasses flipping labels. The parameters $\alpha_1 = 0.2$ and $\alpha_2 = 0.1$ represent the trade-off among pairwise ranking loss, regression loss, and classification loss. Note that, with more crowd labels used in regression, the performance benefited from regression is greater. To make full use of the supervision from the counting anchor set, we can utilize these few anchor samples with count labels as D for regression. In the most extreme case, we can attach all the training images with count labels for optimizing the regression loss term, so that it transforms our formulation into a label-based optimization completely. Yet, with the involvement of the ranking loss, the performance is significantly superior to the pure regression-based methods.

4. Experiments

We conduct extensive experiments to evaluate our approach on several crowd counting benchmarks: ShanghaiTech PartA [59], UCF-QNRF [14], UCF_CC_50 [13]. We compare our approach against other weakly-supervised counting methods. Note that, compared to the existing weakly-supervised methods, our ranking-based model requires weaker supervision. In this section, we first describe the implementation details and evaluation metrics. Then, we compare and evaluate our method with the peer weakly-supervised state-of-the-art methods. Last, we perform comprehensive ablation studies to delve into our model.

4.1. Implementation Details and Metrics

Implementation Details. The network backbone used in our experiment is PVTv2 [52]. Apparently, the backbone can also be replaced by other Transformer-based models or CNN-based models. The proposed network is trained using Adam solver [16] as the optimizer with a mini-batch size of 1. The learning rate is set to 1e-5. Except for the ablation study, the margin of SVM in our methods is set to 0.5. All images are resized to 1152×768.

Evaluation Metrics. There are two metrics widely used to evaluate the performance of crowd counting. Mean Absolute Error (MAE) implies count estimation accuracy, which is formally defined as,

$$MAE = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} |c_i - c_i^{GT}|, \tag{5}$$

where $\mathcal{N}$ is the number of images in the testing set. c_i is the predicted count for i-th image, while its actual count is c_i^{GT}. Mean Squared Error (MSE) is the metric for the variance of counting estimation to reflect the robustness of prediction, which is defined as,

$$MSE = \sqrt{\frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} (c_i - c_i^{GT})^2}. \tag{6}$$

Anchoring Mapping. During inference, to evaluate the proposed ranking method, the scaling mapping function $S : v \rightarrow y$ from potential scores to real count numbers can be learned by linearly fitting the images in the counting anchor set, where $\{v, y\}$ is paired and known.

4.2. Comparison with State-of-the-arts

Compared with previous weakly supervised counting methods, our supervision scheme is unique, and requires less supervision information. Here, we mainly compare the proposed methods with other approaches with diverse supervised settings. Our comparison evaluation is conducted on ShanghaiTech PartA [59] and UCF-QNRF [14] datasets.

ShanghaiTech Part A Dataset. ShanghaiTech Part A dataset [59] is a large-scale crowd counting dataset, which is composed of 482 images with 244,167 annotated persons. The training set includes 300 images with 162,707 annotated persons, and the remaining 182 images are for testing. The images are captured from the Internet and the number of humans ranges from 33 to 3139 per image. Following the assumption of glance annotation, the available training ranking pairs are 24,386. Due to our transitive automatic labeling, the time of manually annotating ranking pairs is reduced to 16,194. We randomly pick up 50 images from the same dataset to set up the counting anchor set, which corresponds to different crowd density levels in the training set. In total, the number of ranking pairs is only $1/10$ of location-based labels. Adding the labeling efforts of counting anchor set, the total labeling amount is around 1/4 of the original labels on the ShanghaiTech Part A dataset.

UCF-QNRF Dataset. The UCF-QNRF dataset [14] contains 1,535 images with counts varying from 49 to 12865 including 1,251,642 annotated heads, thus the average count is around 815 per image. The training set includes 1,201 images, and 334 images are for testing. We also randomly pick up 50 images from different ranges of density as the counting anchor set. Owing to the huge data size, the number of available ranking pairs is exploded, thus it is intractable to annotate all pairs in the real world. To handle the problem, the hyper-parameter $\mathcal{N}_{sim}$ is provided to simulate the number of annotated ranking pairs when training for ranking. Here we set $\mathcal{N}_{sim} = 48,000$, which means the given set P with size 48,000 is fixed before training and these pairs are available training samples.

Categories of Crowd Counting Methods. Generally, the supervision of the evaluation methods can be roughly categorized from laborious to effortless as four levels, location level, count level, ranking level, and no label:

- *Location level* supervision relies on location-based density maps as the optimizing objective. For a dense-scene crowd image, it requires a lot of effort to complete the annotation of hundreds of locations.
- *Count level* supervision is based on crowd count num-

Table 1: Comparison of our proposed method with baselines and related methods on ShanghaiTech Part A [59] and UCF-QNRF [14]. "label level" refers to the supervision level of training. ✔ means the model employs all the labels under the corresponding level of supervision, and ✦ means the model employs a few labels at this supervision level. $*$ indicates the 0.1% of the parameters are tuned with location-level supervision. Note that, *Ours* exploits the same amount of count labels as other weakly supervised methods, and ranking labels can be auto-generated from count labels without extra annotating effort. ✯ indicates our model is purely trained with ranking labels. Best weakly-supervised results are highlighted in **red**.

	Label level			**ST PartA**[59]		**UCF-QNRF**[14]	
Method	**Location**	**Count**	**No label**	**MAE**↓	**MSE**↓	**MAE**↓	**MSE**↓
Fully Supervised Methods							
MCNN [59] (2016)	✔			110.2	173.2	277.0	426.0
Switching-CNN [38] (2017)	✔			90.4	135.0	228.0	445.0
CSRNet [19] (2018)	✔			68.2	115.0	119.2	211.4
CAN [27] (2019)	✔			62.3	100.0	107.0	183.0
ADSCNet [3] (2020)	✔			55.4	97.7	71.3	132.5
TopoCount [1] (2021)	✔			56.9	95.2	87.3	142.4
P2PNet [44] (2021)	✔			52.7	85.1	85.3	154.5
CLTR [22] (2022)	✔			61.2	104.6	89.0	159.0
MAN [23] (2022)	✔			56.8	90.3	77.3	153.5
Weakly-/Semi-/Un-supervised Methods							
GWTA-CCNN [37] (2019)	$*$		✔	154.7	229.4	-	-
CSS-CCNN [35] (2020)			✔	207.3	310.1	442.4	721.6
IRAST (Label Only) [29] (2020)	✦			98.3	159.2	147.7	253.1
IRAST [29] (2020)	✦		✔	86.9	148.9	135.6	233.4
CCLS [56] (2020)		✔		104.6	145.2	-	-
MATT [17] (2021)	✦	✔		80.1	129.4	-	-
TransCrowd [21](2022)		✔		66.1	105.1	97.2	170.3
CrowdMLP [50](2022)		✔		57.8	84.4	94.1	170.3
RFSNet [33](2023)		✔		62.9	93.7	97.9	171.3
CrowdFormer [39](2023)		✔		61.2	93.3	**92.8**	**165.9**
HACC [18](2023)		✔		58.3	84.6	92.9	168.7
Ours (Ranking Only)		✯		71.1	107.3	99.7	177.6
Ours (Partial Weak Labels)		✦		68.2	106.5	98.4	174.3
Ours		✔		**53.4**	**84.4**	**92.3**	**164.2**

bers without location supervision. There are not many labels but one-by-one manual counting is required.

- *No label* supervision is not to use any annotated label, only use raw crowd images for input.

Note that our method can be trained with pure ranking labels for pair-wise images, and these ranking labels can be easily obtained from count labels without additional cost to enhance our weakly supervised training.

Experimental Results. In our experiments, we trained our model using the same amount of weak count labels with other weakly supervised methods. Furthermore, we introduced two variants: one trained solely on ranking labels and another on a combination of ranking and partial count labels (50 samples). These variants highlight the gains achieved with minimal weak samples. In the context of partial count labels, these weak labels can function directly as anchors without necessitating additional fine-tuning.

The quantitative comparison with the state-of-the-art methods on these two datasets is presented in Table 1. The fully-supervised location-level methods are listed in the first part of the table, and the weakly-/semi-/un-supervised methods for comparison are listed in the second part. The performance of the proposed method and compared baselines are at the bottom.

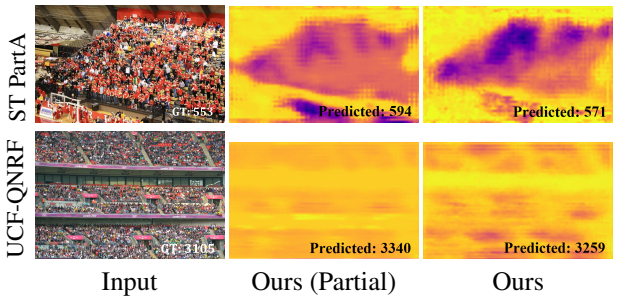

Figure 3: Qualitative visualization of predicted density maps on two examples from ShanghaiTech PartA and UCF-QNRF by our methods. It is shown that ranking-based methods are quite good at distinguishing the crowd regions.

As shown in Table 1, we can see that our model trained with ranking only shows promising results compared to other methods. This demonstrates that these flexible and achievable annotations are practical in crowd counting.

Table 2: Contribution of our proposed method on the ShanghaiTech Part A and UCF-QNRF.

	ST PartA		UCF-QNRF	
Method	**MAE**	**MSE**	**MAE**	**MSE**
Baseline	66.8	119.3	101.4	198.6
Baseline + Multi-Scale	59.5	86.2	93.9	170.1
Baseline + Multi-Scale + MLP-Branch	**53.4**	**84.4**	**92.3**	**164.2**

The additional 50 samples can boost the counting accuracy. When compared with other weakly-supervised methods such as CrowdMLP [50] and HACC [18] with the same amount of weak training data, the performance of *Ours* is the best in the weakly-supervised setting on MAE and MSE, and even close to location level baseline MAN [23]. We visualize the density maps delivered from the first layer of the network as illustrated in Fig. 3. Although there are no location points as supervisory signals, the estimation of our approach is close to the ground truth density maps.

4.3. Ablation Study

Contribution of the multi-scale features. In order to evaluate the role of multi-scale feature fusion in crowd counting, as a baseline, only use the largest size features as input to the next stage. As shown in Table 2, the performance exhibits enhancement as a result of the aggregation of features derived from distinct stages.

Contribution of the Upside-Down MLP. While intuitively assessing whether an image has upside-down flipping as an additional image-level auxiliary classification task appears to be effective, experimental evaluation is necessary to determine whether this auxiliary task can truly assist in improving the model's performance. The results presented in Table 2 indicate that MAE and MSE have improved scores on different datasets. This implies that the additional supervision information we introduce indeed assists the model in comprehending scenarios characterized by diverse distributions.

Impact of the margin on SVM. We set hyper-parameter margin $\mathcal{M}$ to 0.5 in the previous experiments. To investigate the impact of margin, we conduct experiments on ShanghaiTech part A datasets. As shown in Table 3 (left), the results demonstrate setting different margins of SVM does not affect the counting performance significantly. The crowd counting performance is similar to our current results when the margin is set properly, which means the proposed method is robust.

The size of the counting anchor set. So far, we have conducted our experiment with the size of the counting anchor set 50 for *Ours* with partial counting labels and 300 for *Ours*. To investigate the effect of anchor set size, we conduct an experiment that applies different sizes on ShanghaiTech Part A. The results are shown in Table 3 (right), we can observe an improved performance by expanding the set. It implies that the counting method based on regression labels can be incorporated into our setting and effectively boost crowd counting.

Table 3: Impacts of the margin $\mathcal{M}$ (left), and the size of the counting anchor set (right). Note that two factors are varied independently. The best results are highlighted in bold.

$\mathcal{M}$	MAE	MSE	**Set size**	MAE	MSE
0	70.7	107.1	$\lvert B\rvert = 10$	84.2	123.8
0.1	71.3	108.6	$\lvert B\rvert = 30$	71.9	108.4
0.5	**68.2**	106.5	$\lvert B\rvert = 50$	68.2	106.5
1.0	68.7	**103.0**	$\lvert B\rvert = 80$	65.8	101.3
3.0	70.9	109.1	$\lvert B\rvert = 150$	**61.3**	**94.1**

Table 4: Cross dataset experiments on the ShanghaiTech Part A, UCF-QNRF, and UCF_CC_50 datasets for demonstrating the generalization of different methods.

	ST PartA → UCF-QNRF		UCF-QNRF → ST PartA		ST PartA → UCF_CC_50	
Method	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
MCNN [59]	-	-	-	-	397.7	624.1
L2R [28]	-	-	-	-	337.6	434.3
SPN [54]	236.3	428.4	87.9	126.3	368.3	588.4
TransCrowd [21]	-	-	78.7	122.5	-	-
CrowdFormer [39]	162.7	333.9	73.0	121.5	-	-
Ours (Partial Labels)	161.4	334.6	76.0	123.1	322.4	431.5
Ours	152.2	324.9	69.0	116.2	286.8	403.1

4.4. Challenging Experiments

Cross Datasets. We conduct experiments to demonstrate the generalizability of our method across different data domains. The model is trained on one dataset of a source domain and evaluated on another dataset as a target domain. The results are demonstrated in Table 4. Thanks to the Upside-Down MLP, our model demonstrates the robustness of scenes with diverse population distributions, We can observe that the proposed method generalizes well to the unseen evaluation datasets. Especially, the proposed method can be comparable to or even better than the location-level supervised methods.

5. Conclusion

In this paper, we propose a novel weakly-supervised setting, in which we leverage the binary ranking of two images with high-contrast crowd counts as training guidance. In particular, we tailor a Siamese Ranking Network that predicts the potential scores of two images indicating the ordering of the counts. Hence, the ultimate goal is to assign appropriate potentials for all the crowd images to ensure their orderings obey the ranking labels, and then map them to actual crowd counts.

Acknowledgment. This project is supported by the Guangdong Natural Science Funds for Distinguished Young Scholars (No. 2023B1515020097); Singapore Ministry of Education Academic Research Fund Tier 1 (MSS23C002).

References

[1] Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. Localization in the crowd with topological constraints. In *AAAI*, volume 35, pages 872–881, 2021.

[2] Carlos Arteta, Victor Lempitsky, J Alison Noble, and Andrew Zisserman. Interactive object counting. In *ECCV*, pages 504–518, 2014.

[3] Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. Adaptive dilated network with self-correction supervision for counting. In *CVPR*, pages 4594–4603, 2020.

[4] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96, 2005.

[5] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *ECCV*, pages 734–750, 2018.

[6] Xinya Chen, Yanrui Bin, Nong Sang, and Changxin Gao. Scale pyramid network for crowd counting. In *WACV*, pages 1941–1950, 2019.

[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, pages 539–546, 2005.

[8] Diptodip Deb and Jonathan Ventura. An aggregated multicolumn dilated convolution network for perspective-free counting. In *CVPR Workshops*, pages 195–204, 2018.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Luca Fiaschi, Ullrich Köthe, Rahul Nair, and Fred A Hamprecht. Learning to count with regression forest and structured labels. In *ICPR*, pages 2685–2688, 2012.

[11] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov):933–969, 2003.

[12] Ralf Herbrich. Large margin rank boundaries for ordinal regression. *Advances in large margin classifiers*, pages 115–132, 2000.

[13] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *CVPR*, pages 2547–2554, 2013.

[14] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *ECCV*, pages 532–546, 2018.

[15] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David Doermann, and Ling Shao. Crowd counting and density estimation by trellis encoder-decoder networks. In *CVPR*, pages 6133–6142, 2019.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, pages 1–13, 2014.

[17] Yinjie Lei, Yan Liu, Pingping Zhang, and Lingqiao Liu. Towards using count-level weak supervision for crowd counting. *Pattern Recognition*, 109:107616, 2021.

[18] Bo Li, Yong Zhang, Chengyang Zhang, Xinglin Piao, and Baocai Yin. Hypergraph association weakly supervised crowd counting. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023.

[19] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, pages 1091–1100, 2018.

[20] Dongze Lian, Jing Li, Jia Zheng, Weixin Luo, and Shenghua Gao. Density map regression guided detection network for rgb-d crowd counting and localization. In *CVPR*, pages 1821–1830, 2019.

[21] Dingkang Liang, Xiwu Chen, Wei Xu, Yu Zhou, and Xiang Bai. Transcrowd: weakly-supervised crowd counting with transformers. *Science China Information Sciences*, 65(6):160104, 2022.

[22] Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *ECCV*, pages 38–54. Springer, 2022.

[23] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multifaceted attention. In *CVPR*, pages 19628–19637, 2022.

[24] Chenchen Liu, Xinyu Weng, and Yadong Mu. Recurrent attentive zooming for joint crowd counting and precise localization. In *CVPR*, pages 1217–1226, 2019.

[25] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *CVPR*, pages 5197–5206, 2018.

[26] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In *CVPR*, pages 3225–3234, 2019.

[27] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *CVPR*, pages 5099–5108, 2019.

[28] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *CVPR*, pages 7661–7669, 2018.

[29] Yan Liu, Lingqiao Liu, Peng Wang, Pingping Zhang, and Yinjie Lei. Semi-supervised crowd counting via self-training on surrogate tasks. In *ECCV*, pages 242–259, 2020.

[30] Yuting Liu, Miaojing Shi, Qijun Zhao, and Xiaofang Wang. Point in, box out: Beyond counting persons in crowds. In *CVPR*, pages 6469–6478, 2019.

[31] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *ICCV*, pages 6142–6151, 2019.

[32] Yanda Meng, Hongrun Zhang, Yitian Zhao, Xiaoyun Yang, Xuesheng Qian, Xiaowei Huang, and Yalin Zheng. Spatial uncertainty-aware semi-supervised crowd counting. In *ICCV*, pages 15549–15559, 2021.

[33] Zhuangzhuang Miao, Yong Zhang, Xinglin Piao, Yi Chu, and Baocai Yin. Region feature smoothness assumption for weakly semi-supervised crowd counting. *Computer Animation and Virtual Worlds*, page e2173.

[34] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *ECCV*, pages 270–285, 2018.

[35] Deepak Babu Sam, Abhinav Agarwalla, Jimmy Joseph, Vishwanath A Sindagi, R Venkatesh Babu, and Vishal M Patel. Completely self-supervised crowd counting via distribution matching. *arXiv preprint arXiv:2009.06420*, 2020.

[36] Deepak Babu Sam and R Venkatesh Babu. Top-down feedback for crowd counting convolutional neural network. *arXiv preprint arXiv:1807.08881*, 2018.

[37] Deepak Babu Sam, Neeraj N Sajjan, Himanshu Maurya, and R Venkatesh Babu. Almost unsupervised learning for dense crowd counting. In *AAAI*, volume 33, pages 8868–8875, 2019.

[38] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *CVPR*, pages 4031–4039, 2017.

[39] Siddharth Singh Savner and Vivek Kanhangad. Crowdformer: Weakly-supervised crowd counting with improved generalizability. *Journal of Visual Communication and Image Representation*, 94:103853, 2023.

[40] D Sculley. Large scale learning to rank. *Advances in Ranking*, page 58.

[41] Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. Revisiting perspective information for efficient crowd counting. In *CVPR*, pages 7279–7288, 2019.

[42] Zenglin Shi, Pascal Mettes, and Cees GM Snoek. Counting with focus for free. In *CVPR*, pages 4200–4209, 2019.

[43] Vishwanath A Sindagi and Vishal M Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *ICCV*, pages 1861–1870, 2017.

[44] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *ICCV*, pages 3365–3374, October 2021.

[45] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *CVPR*, pages 2325–2333, 2016.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[47] Matthias von Borstel, Melih Kandemir, Philip Schmidt, Madhavi K Rao, Kumar Rajamani, and Fred A Hamprecht. Gaussian process density counting from weak supervision. In *ECCV*, pages 365–380. Springer, 2016.

[48] Elad Walach and Lior Wolf. Learning to count with cnn boosting. In *ECCV*, pages 660–676. Springer, 2016.

[49] Jia Wan and Antoni Chan. Adaptive density map generation for crowd counting. In *ICCV*, pages 1130–1139, 2019.

[50] Mingjie Wang, Jun Zhou, Hao Cai, and Minglun Gong. Crowdmlp: Weakly-supervised crowd counting via multi-granularity mlp. *arXiv preprint arXiv:2203.08219*, 2022.

[51] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *CVPR*, pages 8198–8207, 2019.

[52] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[53] Feng Xiong, Xingjian Shi, and Dit-Yan Yeung. Spatiotemporal modeling for crowd counting in videos. In *ICCV*, pages 5151–5159, 2017.

[54] Chenfeng Xu, Kai Qiu, Jianlong Fu, Song Bai, Yongchao Xu, and Xiang Bai. Learn to scale: Generating multipolar normalized density maps for crowd counting. In *ICCV*, pages 8382–8390, 2019.

[55] Zhaoyi Yan, Yuchen Yuan, Wangmeng Zuo, Xiao Tan, Yezhen Wang, Shilei Wen, and Errui Ding. Perspective-guided convolution networks for crowd counting. In *ICCV*, pages 952–961, 2019.

[56] Yifan Yang, Guorong Li, Zhe Wu, Li Su, and Qingming Huang. Weakly-supervised crowd counting learns from sorting rather than locations. In *ECCV*. Springer International Publishing, 2020.

[57] Anran Zhang, Jiayi Shen, Zehao Xiao, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. Relational attention network for crowd counting. In *ICCV*, pages 6788–6797, 2019.

[58] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *CVPR*, pages 833–841, 2015.

[59] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pages 589–597, 2016.

[60] Muming Zhao, Jian Zhang, Chongyang Zhang, and Wenjun Zhang. Leveraging heterogeneous auxiliary tasks to assist crowd counting. In *CVPR*, pages 12736–12745, 2019.