Deep Rank-Consistent Pyramid Model for Enhanced Crowd Counting

Jiaqi Gao, Member, IEEE, Zhizhong Huang, Graduate Student Member, IEEE, Yiming Lei, Member, IEEE, Hongming Shan, Senior Member, IEEE, James Z. Wang, Senior Member, IEEE, Fei-Yue Wang, Fellow, IEEE, and Junping Zhang, Senior Member, IEEE

Abstract—Most conventional crowd counting methods utilize a fully-supervised learning framework to establish a mapping between scene images and crowd density maps. They usually rely on a large quantity of costly and time-intensive pixel-level annotations for training supervision. One way to mitigate the intensive labeling effort and improve counting accuracy is to leverage large amounts of unlabeled images. This is attributed to the inherent self-structural information and rank consistency within a single image, offering additional qualitative relation supervision during training. Contrary to earlier methods that utilized the rank relations at the original image level, we explore such rank-consistency relation within the latent feature spaces. This approach enables the incorporation of numerous pyramid partial orders, strengthening the model representation capability. A notable advantage is that it can also increase the utilization ratio of unlabeled samples. Specifically, we propose a Deep Rank-consistent Pyramid Model (DREAM), which makes full use of rank consistency across coarse-to-fine pyramid features in latent spaces for enhanced crowd counting with massive unlabeled images. In addition, we have collected a new unlabeled crowd counting dataset, FUDAN-UCC, comprising 4000 images for training purposes. Extensive experiments on four benchmark datasets, namely UCF-QNRF, ShanghaiTech PartA and PartB, and UCF-CC-50, show the effectiveness of our method compared with previous semi-supervised methods. The codes are available at https://github.com/bridgeqiqi/DREAM.

Index Terms—Crowd counting, feature pyramid, ranking, semi-supervised learning.

I. INTRODUCTION

CROWD counting has broad applications in traffic management, public safety surveillance, and smart city planning such as preventing stampedes and estimating participation in rallies or parades. During health crises, such as the COVID-19 pandemic, effective crowd counting can help authorities determine whether social distancing measures are feasible in specific public areas. The primary goal of crowd counting is to estimate the number of people present in a given image or video sequence, especially in crowded scenes. Although crowd counting has been an active research area in recent years, the task remains a challenge due to the influence of many extrinsic factors, including occlusion, illumination, head size variations, diverse perspectives, and uneven distribution of crowds. It may also suffer from significant costs in terms of time and resources in labeling dense scenes.

Earlier counting methods were based on detection techniques [1], [2]. They mainly focused on designing robust human-pose or human-body detectors to count pedestrians in scenes using a sliding-window template matching approach [2]. This process is both computationally expensive and time-consuming. Meanwhile, its accuracy to a great extent depends on detector performance, which may suffer from the occlusion, illumination, and scale variation issues. As an alternative, some researchers treated crowd counting as a regression task [3], [4], [5] by learning a mapping from the original image to the final count number, which can accelerate both the training and inference process.

Both paradigms have shown proficiency in sparse scenes. Even with accuracy improvements by several state-of-the-art object detection methods [6], [7], [8], [9], [10], [11], [12], [13] with the help of extracting multiscale features, performance can be hindered due to occlusion [14], [15], congested scenes [16] and tiny head sizes for individuals situated far from the camera’s perspective. Benefiting from the strong representation learning ability of convolutional neural networks (CNNs), in recent years, CNN-based methods [17], [18], [19],
Surrogate tasks are used to generate pseudo-labels for semi-supervised crowd counting. Specifically, auxiliary age-limited labeled images and abundant unlabeled images must be handled. Other solutions are to leverage the characters in the game and pedestrians in the real world datasets. However, quite a few physical differences between them demand and reducing labeling costs. Synthetic images generated by game environments may serve as a potential solution to address the labeled data shortage in the real world. Furthermore, the labor-intensive and expensive labeling process results in vast amounts of data remains.

Our main contributions can be summarized as follows.

1) We propose a semi-supervised crowd counting framework that exploits the rank-consistent pyramid features among unlabeled images to enhance counting accuracy. With the help of the novel coarse-to-fine feature margin ranking loss, the model could utilize partial orders and rank-consistent information among unlabeled images to estimate the counts more accurately with limited labeled images. The proposed coarse-to-fine ranking loss on the feature level is both intuitive and straightforward to implement.

2) We have constructed a large unlabeled crowd counting dataset, FUDAN-UCC, from the Internet. With 4000 high-resolution images of densely populated scenes, the dataset paves the way for robust, unbiased comparisons among semi-supervised crowd counting methods.

[1] https://www.gettyimages.com/
3) Extensive experiments on several benchmark crowd counting datasets show the effectiveness of our proposed DREAM model.

II. RELATED WORK

A. Conventional Counting Methods

1) Detection-Based Methods: Using object detection for crowd counting [51] is an intuitive approach. In earlier studies, detectors were trained using classical hand-crafted features such as scale-invariant feature transform (SIFT), histogram of oriented gradients (HOG), and edge features extracted from the part or whole of human bodies. Researchers [1], [52], [53], [54] extracted the general features from the entire body to train classifiers using algorithms such as support vector machine (SVM), boosting, and random forest. However, they achieved limited performance, especially in scenes with significant occlusions. In contrast, body-part features [55], [56], e.g., heads and shoulders, improved the accuracy to some extent. Nevertheless, counting-by-detection methods primarily excel in sparse scenes, given their sensitivity to heavy occlusions and density variations.

2) Regression-Based Methods: Regression-based approaches focus on enhancing the capacity to estimate global counts in crowd counting. They typically seek to learn a mapping function from both local and global features to the overall count [3], [5], [57]. The process generally comprises two steps: 1) extracting useful features, including foreground features, textures, corners, HoG, and local binary patterns (LBPs) and 2) Training a regression model such as linear regression, ridge regression, Bayesian Poisson regression [3], and Gaussian process regression [48] based on features extracted from step 1).

Nevertheless, these conventional counting approaches mainly rely on hand-crafted features and may falter in extremely crowded scenes. Meanwhile, by overlooking pedestrian distribution, their performance remains constrained. To better learn spatial distribution information of individuals within a scene, Lempitsky and Zisserman [58] proposed to predict a density map instead of regressing a scalar for crowd counting. A density map can reflect the distribution of people approximately and its integral is equal to the number of people in a given image.

B. CNN-Based Methods

1) Supervised Methods: Most off-the-shelf learning-based methods are built upon stacks of convolutional operations to regress crowd counts or density maps. One density map could not only predict the potential location of each person but also explain the overview of spatial distribution. More specifically, Wang et al. [59] used an AlexNet-like architecture to predict the number of people in highly crowded scenes. Considering the perspective information, Zhang et al. [17] achieved better counting for unseen images in cross-scene scenarios. Zhang et al. [18] further proposed a multicoloum convolutional neural network (MCNN) architecture, which contains multicoloumn convolutional layers with different kernel sizes, to resolve the scale variation issue for crowd counting. Following this, several multicoloumn models were developed. Sam et al. [20] designed a switchable network for training patches within different density levels. Sindagi and Patel [19] introduced both local and global contextual information to assist the network in generating high-quality density maps. Observing that features extracted from different columns normally have similar and redundant patterns, Li et al. [23] proposed a deeper, single-column network within dilated convolutions to address the issues and achieve better performance. Jiang et al. [28] fused features from different layers by simple concatenations in CNN to obtain a multiscale feature map representation. Besides, Liu et al. [24] combined scale-aware contextual features with perspective maps to estimate density maps, while Cao et al. [26] and Chen et al. [27] designed a scale aggregation network and a scale pyramid network to tackle the scale variation problems. Chen et al. [60] proposed a multiscale spatial guided perception aggregation network (MGANet) to deal with the dramatic scale variation issues in a single image. Yang et al. [62] uniformly warped the input images to normalize head sizes at different locations to the same scale through perspective transformations. Furthermore, to correct small errors of ground truth caused by the empirically chosen parameter of head scale \( \sigma \), Wan and Chan [63] and Cao et al. [64] utilized the kernel-based density map to refine the final density map. Zhu et al. [65] designed an end-to-end confusion region discriminating and erasing network (CDENet) to address the incorrect estimation problems among confusion regions. Bai et al. [66] self-corrected the density map by expectation-maximization (EM) algorithm.

Adversarial networks [67], [68] have also been used in crowd counting to generate high-quality density maps. Sindagi and Patel [69] fused multilevel bottom-top and top-bottom features to address the scale variation problems in crowd counting. Xiong et al. [70] divided feature maps into several grids and counted them hierarchically. A few deep reinforcement learning approaches [71], [72] have also been introduced to enhance the crowd counting task. For example, Lu et al. [71] dealt with crowd counting from a sequence decision-making perspective, i.e., weighing densities. Ma et al. [73] proposed a Bayesian loss to learn an expectation of people distribution by using point supervision instead of generating density maps. To estimate the changes of head sizes in a single image more accurately, Lian et al. [74], [75] leveraged a depth prior information and designed a depth-adaptive kernel to generate high-fidelity ground truth density maps for better training.

2) Semi-Supervised/Weakly-Supervised Methods: Fully-supervised counting methods usually require massive pixel-level annotations, which can be prohibitively costly. A synthetic dataset [45] constructed by the GTA5 game environment may solve the data-hungry issue. Another way is to leverage the vast repositories of unlabeled crowd images to assist the model in learning task-specific representations for...
more accurate crowd counting. For instance, Loy et al. [76] implemented a unified active and semi-supervised regression framework to exploit the manifold structure of images. There are two approaches in semi-supervised settings.

1) Generating a set of reliable pseudo labels [77] for unlabeled images and then tune the model in a supervised way. Sindagi et al. [48] and Liu et al. [49] employed the Gaussian process method and surrogate segmentation tasks, respectively, to generate the pseudo-labels for unlabeled images.

2) Exploiting self-structural information and constructing an unsupervised loss among unlabeled images as an auxiliary loss to optimize the model. Liu et al. [47] leveraged more unlabeled data collected from the Internet and constructed a rank margin loss to optimize the model. Sam et al. [78] constructed an unsupervised reconstruction loss to learn useful features from unlabeled images and then trained the counter using labeled images. Parameters in the front layers are frozen when training the subsequent layers.

Utilizing rank-consistent information among unlabeled images is crucial when counting is based on limited labeled samples. Although Liu et al. [47] claimed that such ordinal relations among unlabeled images in the input space are effective, these constraints are actually insufficient. In this article, we explore such relations at different feature levels because features in deep layers of CNN contain more semantic information that is closer to densities and distributions. By the definition of receptive field, relative positions remain consistent through stacks of convolution operations. In other words, a feature patch within an intermediate feature map is supposed to correspond to a specific subregion of the given input image (see Fig. 1). Therefore, the output predicted by one feature patch, $g(FV_2)$, should be no greater than that of its subpatch, $g(FV_1)$. We regard these relations as “partial orders.”

In our research, we exploit these partial orders across multiple intermediate layers and construct a coarse-to-fine feature pyramid margin rank loss to assist the model in learning from unlabeled images. In this case, our proposed DREAM model can increase the utilization ratio of unlabeled images which are at least three times that of L2R [47].

C. Learning to Rank

Learning to rank [79] aims to rank the items according to their relevance to a given query, which is a critical research topic in many applications, including information retrieval [80], [81], recommender systems [82], [83], [84], confidence estimation [85], image retrieval [86], [87], [88], and image quality assessment [89], [90], [91]. In information retrieval [80], [81], a ranking function computes and assigns scores to texts, documents, or images, and then sorts them in descending order to facilitate retrieval. In computer vision tasks, Zagoruyko and Komodakis [89] analyzed the feature similarity with different CNN architectures by comparing the image patches. Faigenbaum-Golovin and Shimshi [90] ranked the pairwise image distortion level to assess the image quality. Additionally, Liu et al. [91] used learning from rankings as a data augmentation technique to train a large neural network that is prepared for fine-tuning.

Our approach has a similar idea to the “learning to rank” methodology, specifically in the ranking of different feature pairs. However, there are two main distinctions. First, our ranking process is unsupervised; there are no human annotations during the ranking process, which contrasts with the common supervised learning approach used in learning to rank. Second, our method ranks different feature pairs with the human densities through an intuitive concept of rank consistency between the subregions of features, instead of relying on a relevance score in learning to rank.

III. METHODS

A. Motivation

In the area of crowd counting, a common sense is that for any image of arbitrary size, the number of people in an image patch is always greater than or at least equal to the number in its subregions [47]. Inspired by this observation and the definition of the receptive field, we believe this common sense is also valid at the feature level. This is because convolution and pooling layers in CNNs maintain the relative positions of the objects in an image. In other words, each position in the intermediate feature maps should represent the specific corresponding regions in the input space, which is often called the receptive field. Thus, it should be guaranteed that the output counts predicted by larger feature patches in the latent space are no fewer than the counts predicted by their subpatches in the same feature map, as shown in Fig. 1. Additionally, this consistency in partial orders is expected across different intermediate layers of the network. In this way, we can greatly increase the utilization ratio of partial orders and structural information among unlabeled images. Further, we visualize the receptive field, feature patches in hidden layers, and the prediction and ground truth of our proposed method in Fig. 2. The visualization results clearly validate our hypothesis that partial orders persist in the feature layers.
B. Feature Pairs Set Generation for Unlabeled Images

To compute the margin rank loss of unlabeled images, a prerequisite is to obtain a set of feature patch pairs satisfying the rank consistency. We generate the rank-consistent feature pairs set from low-level to high-level layers. For any given feature map $F \in \mathbb{R}^{C \times H \times W}$ within a given layer, we crop $M$ subregions, all sharing the same center point and with exponentially decreased cropped ratio $r^M$, where $r$ ranges between 0 and 1. The center point is randomly sampled from a small region centered in the feature map, whose size is $(1/8)H \times (1/8)W$. After the cropping process, specifically, we guarantee that every subregion is totally contained by its larger subregion for training. Together with the feature map $F$ itself, we can obtain $M + 1$ feature patches, with any two forming potential candidates for the rank-consistent feature pair set $S$, within a layer. Formally, we have

$$S = \{(v_m, v_n) | m < n, v_m \cap v_n = v_m, v_m \cup v_n = v_n\}$$ (1)

where $m, n \in \{0, 1, 2, \ldots, M\}$ and $M$ is set to 4 in practice. $v_i$ represents the cropped feature patches, with $v_0$ being the smallest feature patch and $v_M = F$ the largest. Therefore, each feature patch $v_m$ is wholly contained in feature patches $v_n$, $\forall n > m$. More details of generating feature patch pairs set in one layer can be seen in Algorithm 1.

C. Model Architecture

We describe the network architecture in detail here. Our network mainly consists of two modules, the feature extractor and the crowd density map estimator. The feature extractor learns coarse-to-fine features through several convolutions and max-pooling operations, while the crowd density map estimator regresses the density map based on these features. Our backbone of the feature extractor module is derived from the VGG-16 network [92]. We only use the first ten layers of VGG-16 with pretrained weights to train our feature extractor

$$v^{(i)} = f(x^{(i)}; \theta)$$ (2)

where the feature extractor module $f(\cdot; \theta)$ with the parameters $\theta$ contains the first ten layers of pretrained VGG-16 network. The module learns the mapping function from the $i$th input image $x^{(i)}$ to output its corresponding feature $v^{(i)}$.

The dilated convolutional layers with $3 \times 3$ kernel size, dilated ratio 2, and stride 1 followed by an upsampling layer constitute the density map estimator, same as described by CSRNet-B [23]

$$D^{(i)} = g(v^{(i)}; \phi)$$ (3)

where $D^{(i)}$ are the predicted density map of the $i$th image. $g(\cdot; \phi)$ is the density estimator with the parameters $\phi$ that contains six dilated convolutional layers, $1 \times 1$ conv layer,
and an upsampling layer. We use the pixel-wise $L_2$ loss as our supervised loss, $L_s$, among labeled images

$$L_s = \frac{1}{2N} \sum_{i=1}^{N} \left\| g \left( f(x^{(i)}; \theta); \phi \right) - y^{(i)} \right\|_2^2 \quad (4)$$

where $N$ is the number of training images in a batch, and $x^{(i)}$ and $y^{(i)}$ are the $i$th original input image and corresponding ground truth density map in one batch, respectively.

For training the unlabeled images, we use the same feature extractor architecture with shared parameters shown in Fig. 3.

D. Coarse-to-Fine Feature Pyramid Margin Ranking Loss

In a deep CNN, feature maps are downsampled using convolutional or pooling layers. The receptive field shows that each pixel in the intermediate feature maps captures the information from one region of the input space. As discussed in Section III-A, larger regions contain the same number or more individuals compared to smaller subregions within the input space. Similarly, because of the receptive field, the counts of the predicted density map from smaller regions of feature maps in the latent space should be equal to or less than those derived from super-regions of the same feature maps.

We adopt the margin ranking loss for training the unlabeled images. We expect the network to learn the ordinal relations of those cropped rank-consistent feature patch pairs in the latent space. Meanwhile, we should guarantee that rank consistency is maintained across pyramid features at different latent layers. More specifically, we crop the feature maps among unlabeled images in $K$ latent layers and construct a margin ranking loss. For the $i$th unlabeled image, the loss $L_r^{(i)}$ is defined as follows:

$$L_r^{(i)} = \max \left( 0, g(v^{(i)}_{u,m}) - g(v^{(i)}_{u,n}) + \epsilon \right) \quad (5)$$

where $v^{(i)}_{u,m}$ and $v^{(i)}_{u,n}$ make up a rank-consistent feature patch pair of the $i$th unlabeled image from $S$ in the latent space. The subscript $u$ in $v_u$ denotes that the cropped patch is from an unlabeled image. The subscript $m$ and $n$ represent any two of the corresponding feature pairs. $\epsilon$ is the margin, indicating the model error-tolerance capacity. We expect the densities from a smaller region of feature maps $g(v^{(i)}_{u,m})$ in latent space to be no more than that from its super-region $g(v^{(i)}_{u,n})$. Therefore, when the network predicts the correct ordinal relation $g(v^{(i)}_{u,m}) \leq g(v^{(i)}_{u,n})$, the loss $L_r^{(i)}$ becomes zero, eliminating the need for gradient backpropagation. Otherwise, the loss $L_r^{(i)}$ captures the difference value between these two estimates, leading to gradient backpropagation that updates the model parameters. Consequently, the rank-consistent feature pyramid margin ranking loss among unlabeled images is defined as follows:

$$L = \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{m=0}^{M-1} \sum_{n=m+1}^{M} \max \left( 0, D^{(i)}_{u,m,k} - D^{(i)}_{u,n,k} + \epsilon \right)$$

$$L_u = \frac{2}{NK(M+1)} L \quad (6)$$

where $N$ is the number of unlabeled images we used during training. $K$ is the number of coarse-to-fine latent spaces selected to generate the rank-consistent feature pairs set $S$. $M$ is the number of cropped patches from the same latent space $k$. $D^{(i)}_{u,m,k}$ and $D^{(i)}_{u,n,k}$ are the density maps predicted by the ordinal relation feature pairs $(v^{(i)}_{u,m}, v^{(i)}_{u,n})$ in set $S$ in the $k$th latent space. In our experiments, $M$ was set to 4 and $K$ to 3. Therefore, there are a total of $K$ feature pairs for a single image, and each pair contains $M + 1$ feature regions (the input feature maps are also included), leading to $K \times (M + 1)$ predicted density maps in the margin ranking loss.

We have introduced the rank-consistent feature pairs set generation method including the main architecture of our model, the fully-supervised loss $L_s$ for labeled images, and the rank-consistent margin ranking loss $L_u$ for unlabeled images. Thus, the final training loss we employed is the combination of $L_s$ and $L_u$, adjusted by the hyper-parameter $\lambda$.

$$L_{\text{total}} = L_s + \lambda L_u. \quad (7)$$

IV. DATASETS AND IMPLEMENTATION DETAILS

A. Experimental Setups

Existing semi-supervised crowd counting methods mainly follow two different training settings. First, off-the-shelf benchmark datasets—such as ShanghaiTech PartA [18], ShanghaiTech PartB [18], UCF-CC-50 [42], and UCF-QNRF [43]—are usually partitioned into labeled and unlabeled subsets with different proportions (namely, 5%, 25%, 30%, and 50% of the full dataset marked for labeled images). Second, other extra crowded images as the unlabeled data are collected for semi-supervised training. For a fair comparison, in this article, we conducted experiments under both training settings, strictly following most previous literature. Unless noted otherwise, the 100% labeled data under the semi-supervised mode has used our newly collected unlabeled dataset.

B. Public Labeled Datasets

1) UCF-CC-50 Dataset [42]: The UCF-CC-50 dataset is the first large-scale congested scene dataset for pedestrian counting. Only 50 images with varying resolutions are collected among available images from public websites. An average of 1279 persons appeared in each image, with counts ranging from a maximum of 4543 to a minimum of 94. Both the small number of images and the drastic variations in the number of people pose a considerable challenge for the counting task. We used a fivefold cross-validation approach for testing due to limited samples. We generated the density map as ground truth by geometry-adaptive Gaussian kernels [18] for a fair comparison.

2) ShanghaiTech PartA Dataset [18]: The ShanghaiTech PartA dataset consists of 482 images with varying resolutions. 241677 heads are annotated in different illumination conditions and crowded scenes. The density distributions span a broad spectrum, ranging from 33 to 3139 persons, with an average of 501 persons per image. We used geometry-adaptive kernels to generate the ground truth for all the images. For supervised learning, researchers often split this dataset into two parts for training and testing.
3) ShanghaiTech PartB Dataset [18]: The ShanghaiTech PartB dataset comprises 716 images with 88,488 head annotations captured from busy streets in the central business districts of Shanghai. The images have a fixed resolution of 768 \times 1024 pixels. The ground truth for these images is generated by a fixed Gaussian kernel with a variance \( \sigma \) of 768. The images search engine GettyImages. It covers complex real-world scenarios, different people distribution, and various illumination conditions.

4) UCF-QNRF Dataset [43]: To construct a larger crowd counting dataset that includes a dramatic variation of head sizes, diverse viewpoints and perspectives, and different locations and times of day, 1535 images are collected from several search engines, including Google Image Search and Flickr. Over 1,251,642 coordinates are labeled, costing more than 2000 labor-hours. Due to the high resolution of the images, we limit the shorter side of a given image to a maximum of 1920 pixels to fit the memory capacity. Images are rescaled with the same ratio to refrain from the global and local contextual information loss as little as possible. The training and test sets consist of 1201 and 334 images, respectively. Similar to the ground truth generation method in the ShanghaiTech PartB dataset, a fixed Gaussian kernel is adopted.

C. FUDAN-UCC Unlabeled Dataset

We captured 4000 images in total from the GettyImages search engine using the keyword “crowd” to create the unlabeled dataset for our experiments. Fig. 4 illustrates a few natural images of our collected dataset that contain varying scenarios, diverse people distribution, and different illuminations. Image resolutions range from 221 \times 612 to 612 \times 612. This captured dataset may serve as a potential standard unlabeled crowd counting benchmark dataset for academic researchers to investigate semi-supervised or unsupervised learning methods in the future. The FUDAN-UCC dataset is publicly available, but is strictly intended for academic and noncommercial purposes. Anyone interested in using it should obtain the necessary licenses and comply with applicable laws and regulations before downloading.

D. Implementation Details

We followed similar preprocessing methods in prior research [48], [49] for supervised training. For semi-supervised learning, we randomly chose the training set with different proportions of labeled and unlabeled samples. Considering the images with larger resolutions of the UCF-QNRF dataset and memory constraints, we resized the shorter side to a maximum of 1920 pixels and maintained the original aspect ratio. Effective data augmentations like random horizontal flipping, random cropping, and normalization were also adopted for both labeled and unlabeled images for training. The captured images were also resized with the same aspect ratio to fit the cropping operations. We used the Adam optimizer with the learning rate of 10^{-5} and weight decay of 10^{-4} in all of our experiments. For the rank-consistent feature pairs set generation process, we set \( K \) to 3, i.e., low-level, mid-level, and high-level from the feature extractor, and the cropped ratio \( r \) to 0.75.

V. EXPERIMENTAL RESULTS

A. Evaluation Metrics

We employed two commonly-used evaluation metrics [mean absolute error (MAE) and root mean squared error (RMSE)] to evaluate our model. The formulae of MAE and RMSE are defined as follows:

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{Y}_i - Y_i|, \quad \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{Y}_i - Y_i)^2}
\] (8)

where \( N \) denotes the number of images from the test set. \( \hat{Y}_i \) and \( Y_i \) are the predicted counts and actual counts of the \( i \)th image, respectively. Briefly speaking, MAE implies the precision of the estimates, and RMSE implies the robustness of the estimates. RMSE is more sensitive to the outliers. Obtaining a model with a low MAE as well as a low RMSE is our expectation.

B. Evaluation on the UCF-CC-50 Dataset

The experimental results on the UCF-CC-50 dataset are shown in Table I. Since there are only 50 images in the UCF-CC-50 dataset, it is not suitable for dividing into labeled and unlabeled datasets. We just utilized the unlabeled images from the collected dataset to train our model. For a fair comparison, we used the same fivefold cross-validation to compute the average MAE and average RMSE metrics. The baseline model was trained by using only labeled 50 images in a fully-supervised way. It achieved 266.10 average MAE. We reproduced the L2R method using our model architecture. It improved the performance on this dataset and led to 261.60 average MAE. Our approach thus achieved an MAE improvement of nearly 10 compared with the L2R method.

C. Evaluation on the ShanghaiTech Dataset

We demonstrate the experimental results on both the ShanghaiTech PartA and PartB datasets in Table II, respectively. All labeled and unlabeled images were chosen from the ShanghaiTech datasets. We randomly picked 5%, 25%, and 50% labeled images from the training set as labeled samples, while the rest of the training set were regarded as unlabeled samples for training. We compared our method against five previous methods, L2R [50], Sindagi et al. [48], IRAST [49], and SUA [93], and STC-Crowd [94]. These methods used different ratios of labeled images to train their models and
report their results. For a fair comparison to them, we used the same proportion of labeled images. For PartA, we reached the best results when the ratio was 50% with MAE 78.4. For 5% and 25% settings, we obtained competitive results for MAE and RMSE compared with Sindagi et al. [48]. A possible explanation is that 5% means only 15 labeled images for training, and the ranking loss among the remaining unlabeled images only reflects qualitative relations which cannot help the model predict the specific count number accurately with a small number of labeled images. As for the PartB dataset, we observed that partial orders among unlabeled images are of limited benefit, especially when the number of labeled samples in the training dataset was relatively small. This can be attributed to two primary reasons: 1) the density distribution is relatively sparse in this dataset and the number of people is relatively small and 2) our proposed qualitatively partial orders among unlabeled images may be more suitable and efficient for crowded scenes. Nevertheless, our proposed method still achieved performance comparable to other methods. When we used all images from the training set and our collected unlabeled ones for training, the results of DREAM achieved the lowest MAE and RMSE scores.

Our model achieved the most significant performance improvements with 50% labeled images, compared to 5% or 100%. It is worth emphasizing that although the proposed method does not require human annotations for unlabeled data, it still benefits from the increasing number of labeled training data. Using only 5% labeled data does not provide the model with sufficient supervision signals. On the contrary, increasing the labeled data to 50% (ten times than 5%) significantly improves the performance. The performance gain decreases when further increasing the labeled data to 100% (only two times for 50% to 100%).

### D. Evaluation on the UCF-QNRF Dataset

The experimental results on the UCF-QNRF dataset are shown in Table II. We also randomly chose 5%, 25%, 30%, and 50% labeled images of its training set and regarded the rest as unlabeled ones for semi-supervised training to compare with previous methods [47], [48], [49]. Moreover, we reproduced the L2R [50] method using our collected unlabeled dataset. The results indicate that our proposed DREAM is superior to the previous semi-supervised methods.

### E. Visualization and Analysis

We visualize the results on these four different datasets in Fig. 5. The visualization results clearly demonstrate the effectiveness of our proposed DREAM model. The performance of utilizing the rank consistency among unlabeled images is better than that of training on labeled ones only (Baseline). To be more specific, in the first and second row of Fig. 5, the baseline predicted density map is prone to learn the uniform distribution of crowded people especially in the crowded regions, training on labeled images only. The rest three methods can address this problem by adding more unlabeled data. However, our proposed method can achieve the best performance with better estimated density maps by utilizing the rank consistency at pyramid features.

### F. Ablation Study

1) Different Utilization Ratios of Labeled Images: We designed the ablation study to verify whether our approach would be robust under the different settings of the varying number of labeled images. We conducted this experiment on the ShanghaiTech PartA dataset. We randomly chose 5%, 25%, and 50% images, respectively, to make up the labeled dataset and the rest images from the training set were the unlabeled ones. The unlabeled images from FUDAN-UCC dataset were only used when the utilization ratio was 100%. As shown in Table III, our method achieved a consistent performance improvement toward training with only labeled images.

2) Impact of Varying $\lambda$: Further, to exploit the role of margin ranking loss in the final mixed loss, we tried different hyper-parameters $\lambda$ which represent the weight of self-supervised loss. The value of $\lambda$ was chosen from {0.1, 0.5, 1, 5, 10}. The specific performance with different values of $\lambda$ is shown in Table IV. We used the unlabeled images from our collected dataset together with all images from the training set of the ShanghaiTech PartA dataset for evaluating the impact of $\lambda$. Our model achieved the best performance on the ShanghaiTech PartA dataset when $\lambda$ is set to 1.

3) Combination of Ranking Losses in Different Layers: Coarse-to-fine pyramid features in different layers represent high-level semantic knowledge and low-level visual signals including textures, edges, backgrounds, among others. Moreover, partial orders should exist in feature patches from different layers. Therefore, we report the different results caused by a diverse combination of ranking loss in disparate layers (low-level, mid-level, and high-level), as shown in Table V. We discovered that the performance would be better as the utilization ratio of coarse-to-fine pyramid feature patches with partial orders increases. In Fig. 6, we also visualize the results of the ranking loss used in different layers. We observed that the low level with the worst performance could not capture the semantic information for crowd counting, the middle level produced a uniform density map, and the high level might be the most helpful than the other two feature levels. Combining the three feature levels achieved the best counting performance with high-quality estimated density maps.

---

**Table I**

| Method          | UCF-CC-50 | Mode | Avg. MAE | Avg. RMSE |
|-----------------|-----------|------|----------|-----------|
| Baseline        | Fully     | 266.10 | 397.50   |
| L2R [50]        | Semi      | 261.60 | 368.07   |
| DREAM (Ours)    | Semi      | 251.52 | 341.06   |

---

**Table II**

| Method          | UCF-CC-50 | Mode | Avg. MAE | Avg. RMSE |
|-----------------|-----------|------|----------|-----------|
| Baseline        | Fully     | 266.10 | 397.50   |
| L2R [50]        | Semi      | 261.60 | 368.07   |
| DREAM (Ours)    | Semi      | 251.52 | 341.06   |
TABLE II

| Method          | Labeled images | Mode | ShanghaiTech PartA MAE | ShanghaiTech PartB MAE | UCF-QNRF MAE |
|-----------------|----------------|------|------------------------|------------------------|--------------|
| Baseline        | Fully          |      | 110.0                  | 211.9                  | 4.2          |
| L2R [50]        | Semi           | 5%   | 115.0                  | 208.0                  | 4.1          |
| Sindagi et al. [48] | Semi          | 25%  | 102.0                  | 172.0                  | 4.2          |
| DREAM (Ours)    | Semi           | 30%  | 112.7                  | 165.7                  | 30.3         |
| Baseline        | Fully          |      | 110.0                  | 160.0                  | -            |
| Sindagi et al. [48] | Semi          | 25%  | 91.0                   | 149.0                  | -            |
| DREAM (Ours)    | Semi           | 30%  | 93.5                   | 148.4                  | 14.9         |
| Baseline        | Fully          |      | 98.3                   | 159.2                  | 23.0         |
| L2R [50]        | Semi           | 30%  | 90.3                   | 153.5                  | 24.4         |
| IARAST [49]     | Semi           | 30%  | 86.9                   | 148.9                  | 22.9         |
| DREAM (Ours)    | Semi           | 30%  | 86.5                   | 121.2                  | 23.8         |
| Baseline        | Fully          |      | 102.0                  | 149.0                  | 27.7         |
| Sindagi et al. [48] | Semi          | 50%  | 89.0                   | 148.0                  | 25.1         |
| SUA [93]        | Semi           | 50%  | 68.5                   | 121.9                  | 20.6         |
| DREAM (Ours)    | Semi           | 50%  | 78.4                   | 112.9                  | 15.3         |
| Baseline        | Fully          |      | 69.1                   | 103.0                  | 16.0         |
| SUA [93]        | Fully          | 50%  | 66.9                   | 125.6                  | 17.9         |
| STC-Crowd [94]  | Fully          | 100% | 62.3                   | 103.5                  | 14.6         |
| L2R [50]        | Semi           | 100% | 73.6                   | 106.6                  | 21.4         |
| L2R* [50]       | Semi           | 100% | 64.4                   | 105.0                  | 13.9         |
| DREAM (Ours)    | Semi           | 100% | 62.6                   | 102.0                  | 13.4         |

Fig. 5. Visual comparisons of different competitors including the baseline, SUA [93], L2R [50], and our method. The four images come from UCF-CC-50, ShanghaiTech PartA, UCF-QNRF, and ShanghaiTech PartB datasets, respectively. (a) Image. (b) Ground truth. (c) Baseline. (d) SUA. (e) L2R. (f) Ours.

4) Different Unlabeled Datasets: We conducted an ablation study to show the effectiveness of our proposed FUDAN-UCC as an unlabeled dataset. Specifically, under the same setting in our paper, the ShanghaiTech PartA dataset was adopted as the labeled dataset while the ShanghaiTech PartB, UCF-QNRF, and FUDAN-UCC were employed as unlabeled datasets for...
TABLE V
Combination of Rank Losses in Different Layers on the ShanghaiTech PartA Dataset

| low | mid | high | MAE | RMSE |
|-----|-----|------|-----|------|
| ✓   | ✓   | ✓    | 65.3| 103.8|
| ✓   | ✓   | ✓    | 63.6| 108.4|
| ✓   | ✓   | ✓    | 63.4| 107.0|
| ✓   | ✓   | ✓    | 62.6| 102.0|

Fig. 6. Visualization results of our proposed rank-consistent feature ranking loss applied to different feature levels. (a) Image. (b) Ground truth. (c) Combination. (d) Low. (e) Mid. (f) High.

TABLE VI
Ablation Study of Different Unlabeled Datasets. Baseline denotes that no additional unlabeled datasets are used

| Unlabeled dataset | MAE | RMSE |
|-------------------|-----|------|
| Baseline          | 69.1| 103.0|
| ShanghaiTech PartR| 65.3| 117.7|
| UCF-QNRF          | 68.9| 108.8|
| FUDAN-UCC         | 62.6| 102.0|

5) Margin Ranking Loss on Different Training Data: We conducted the experiments when adding this margin loss to the different combinations of training data in Table VII. Compared to the baseline, the results suggest that the performance can be improved by introducing margin ranking loss. Leveraging more unlabeled images can further boost the performance compared to train on only labeled images. Applying the margin ranking loss to both unlabeled and labeled data does not exhibit significant differences from the ones on only unlabeled data. We argue that it is because ranking consistency between different feature pairs can be easily learned by the counting model when the supervision training signals exist. Consequently, the margin ranking loss hardly contributes to the counting model using the labeled data.

G. Limitations and Discussions

In this article, we introduce a deep rank-consistent pyramid model called DREAM for improving crowd counting especially when the labeled images are limited. According to the analysis of extensive experimental results and visualization comparisons, DREAM can rival state-of-the-art methods. However, it has certain limitations, which we will discuss in this section.

1) Failure Cases: Although DREAM achieves superior estimated counts compared to baseline and other methods (see Fig. 5), we have identified instances where it faltered in Fig. 7: 1) our model cannot distinguish real people and human-like sculptures and 2) our model cannot detect patchy back-side heads, which appear almost completely black.

2) Imperfect Annotations: Due to the costly pixel-wise labeling process, imperfect annotations are bound to occur in public datasets. These include missing or ambiguous annotations, background noises, annotation shifts, and the like. In this article, we mainly focus on the semi-supervised setting for crowd counting without considering the issue with noisy annotations. However, we believe that our method can mitigate such issues to a degree because the proposed rank-consistent model does not use manual annotations among unlabeled images during training. In addition, our model is orthogonal to the current literature. We can integrate ideas from other techniques such as modeling noises [95] and cross-head [96] to address these challenges.

3) FUDAN-UCC Versus Synthetic Datasets: In this study, we collected an unlabeled dataset, the FUDAN-UCC, to assist the crowd counting model training when the labeled data is limited. The data-hungry issue of crowd counting remains challenging because of the time-consuming labeling process, especially for densely populated scenes. To address this issue, some researchers [45], [74] introduced a few virtual datasets with synthetic images to provide sufficient training data. Wang et al. [45] generated a GCC dataset with different people distribution, backgrounds, weather conditions, illumination, and camera perspectives by changing the hyperparameters of the game environment GTA5. Lian et al. [74] provided the ShanghaiTechRGBD-Syn dataset with depth prior information
and designed a depth-adaptive kernel to generate high-fidelity density maps for better training. Unfortunately, the significant domain gap between synthetic and real-world data is inevitably introduced, which makes it challenging to simply train on synthetic data for better performance. Instead of synthesizing training data, our motivation to collect FUDAN-UCC is that, as a real-world dataset, FUDAN-UCC could be directly used for assisting the model training without making any domain transfer. We believe that it could serve as a public real-world unlabeled dataset for comparing future semi-supervised crowd counting methods.

VI. CONCLUSION

Our work focused on taking advantage of partial orders from coarse-to-fine pyramid features to assist the neural network to enhance the qualitative discrimination among unlabeled images. Extensive experiments show that DREAM outperforms other state-of-the-art methods with the help of self-supervised coarse-to-fine feature pyramid ranking loss, especially in dense crowd scenes. Being simple and intuitive, our proposed method is easy to implement. Besides, our new dataset FUDAN-UCC can be a valuable addition to the community that could redefine the benchmark for unlabeled data in semi-supervised crowd counting. We believe the insights gained and potential solutions discussed provide a foundation for future innovations in this space.

REFERENCES

[1] B. Leibe, E. Seemann, and B. Schiele, “Pedestrian detection in crowded scenes,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2005, pp. 878–885.

[2] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: An evaluation of the state of the art,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 4, pp. 743–761, Apr. 2012.

[3] A. B. Chan and N. Vasconcelos, “Bayesian Poisson regression for crowd counting,” in Proc. IEEE 12th Int. Conf. Comput. Vis., Sep. 2009, pp. 545–551.

[4] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, “Crowd counting using multiple local features,” in Proc. Digit. Image Comput., Techn. Appl., Dec. 2009, pp. 81–88.

[5] K. Chen, C. C. Loi, S. Gong, and T. Xiang, “Feature mining for localised crowd counting,” in Proc. Brit. Mach. Vis. Conf., 2012, p. 3.

[6] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 91–99.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.

[8] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, arXiv:1804.02767.

[9] W. Liu et al., “SSD: Single shot MultiBox detector,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 21–37.

[10] L. Jiao et al., “New generation deep learning for video object detection: A survey,” IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 8, pp. 3195–3215, Aug. 2022.

[11] J. Cao, Y. Pang, J. Han, and X. Li, “Hierarchical regression and classification for accurate object detection,” IEEE Trans. Neural Netw. Learning Syst., vol. 34, no. 5, pp. 2425–2439, May 2023.

[12] Z. Wu, J. Wen, Y. Xu, J. Yang, X. Li, and D. Zhang, “Enhanced spatial feature learning for weakly supervised object detection,” IEEE Trans. Neural Netw. Learn. Syst., early access, Jun. 8, 2022, doi: 10.1109/TNNLS.2022.3178180.

[13] C. Li, F. Liu, Z. Tian, S. Du, and Y. Wu, “DAGCN: Dynamic and adaptive graph convolutional network for salient object detection,” IEEE Trans. Neural Netw. Learn. Syst., early access, Nov. 14, 2022, doi: 10.1109/TNNLS.2022.3219245.

[14] H. Tan, X. Liu, B. Yin, and X. Li, “MHSA-Net: Multithead self-attention network for occluded person re-identification,” IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 11, pp. 8210–8224, Nov. 2023.

[15] J. Miao, Y. Wu, and Y. Yang, “Identifying visible parts via pose estimation for occluded person re-identification,” IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 9, pp. 4624–4634, Sep. 2022.

[16] Z. Wang, J. Zhan, C. Duan, X. Guan, P. Lu, and K. Yang, “A review of vehicle detection techniques for intelligent vehicles,” IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 8, pp. 3811–3831, Aug. 2023.

[17] C. Zhang, H. Li, X. Wang, and X. Yang, “Cross-scene crowd counting via deep convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 833–841.

[18] 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. (CVPR), Jun. 2016, pp. 589–597.

[19] V. A. Sindagi and V. M. Patel, “Generating high-quality crowd density maps using contextual pyramid CNNs,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1879–1888.

[20] D. B. Sam, S. Surya, and R. V. Babu, “Switching convolutional neural network for crowd counting,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4031–4039.

[21] L. Boominathan, S. S. S. Kruithveni, and R. V. Babu, “CrowdNet: A deep convolutional network for dense crowd counting,” in Proc. 24th ACM Int. Conf. Multimedia, Oct. 2016, pp. 640–644.

[22] N. Liu, Y. Long, C. Zou, Q. Niu, L. Pan, and H. Wu, “ACD-CrowdNet: An attention-injective deformable convolutional network for crowd understanding,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3220–3229.

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

[24] W. Liu, M. Salzmann, and P. Fua, “Context-aware crowd counting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5094–5103.

[25] Y. Tian, Y. Lei, J. Zhang, and J. Z. Wang, “PaDNNet: Pan-density crowd counting,” IEEE Trans. Image Process., vol. 29, pp. 2714–2727, Aug. 2020.

[26] X. Cao, Z. Wang, Y. Zhao, and F. Su, “Scale aggregation network for accurate and efficient crowd counting,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 734–750.

[27] X. Chen, Y. Bin, N. Sang, and C. Gao, “Scale pyramid network for crowd counting,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2019, pp. 1941–1950.

[28] X. Jiang et al., “Learning multi-level density maps for crowd counting,” IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 8, pp. 2705–2715, Aug. 2020.

[29] L. Dong, H. Zhang, J. Ma, X. Xu, Y. Yang, and Q. M. J. Wu, “CLRNet: A cross locality relation network for crowd counting in videos,” IEEE Trans. Neural Netw. Learn. Syst., early access, Oct. 10, 2022, doi: 10.1109/TNNLS.2022.3209918.

[30] J. Gao, T. Han, Y. Yuan, and Q. Wang, “Domain-adaptive crowd counting via high-quality image translation and density reconstruction,” IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 8, pp. 4803–4815, Aug. 2023.

[31] C. Zhou, C. Xu, Z. Cui, T. Zhang, and J. Yang, “Self-teaching video object segmentation,” IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 4, pp. 1623–1637, Apr. 2022.

[32] G. Li et al., “Self supervised progressive network for high performance video object segmentation,” IEEE Trans. Neural Netw. Learn. Syst., early access, Nov. 16, 2022, doi: 10.1109/TNNLS.2022.3219936.

[33] G. Xian et al., “Location-guided LiDAR-based panoptic segmentation for autonomous driving,” IEEE Trans. Intell. Vehicles, vol. 8, no. 2, pp. 1473–1483, Feb. 2023.

[34] Z. Gu, S. Zhou, L. Niu, Z. Zhao, and L. Zhang, “From pixel to patch: Synthesize context-aware features for zero-shot semantic segmentation,” IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 10, pp. 7689–7703, Oct. 2023.
[78] 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., vol. 33, 2019, pp. 8868–8875.

[79] T.-Y. Liu, “Learning to rank for information retrieval,” Found. Trends® Inf. Retr., vol. 3, no. 3, pp. 225–331, 2009.

[80] H. Li, Learning to Rank for Information Retrieval and Natural Language Processing. Cham, Switzerland: Springer, 2022.

[81] E. Ghanbari and A. Shakerly, “A learning to rank framework based on cross-lingual loss function for cross-lingual information retrieval,” Appl. Intell., vol. 52, no. 3, pp. 3156–3174, 2022.

[82] X. Chen et al., “Set-to-sequence ranking-based concept-aware learning path recommendation,” 2023, arXiv:2306.04234.

[83] M. Zehlike, K. Yang, and J. Stoyanovich, “Fairness in ranking, Part II: Learning-to-rank and recommender systems,” ACM Comput. Surv., vol. 55, no. 6, pp. 1–41, Jul. 2023.

[84] H. Wang, “Skellam rank: Fair learning to rank algorithm based on Poisson process and Skellam distribution for recommender systems,” 2023, arXiv:2306.06607.

[85] C. Li, X. Hu, and C. Chen, “Confidence estimation using unlabeled data,” in Proc. 11th Int. Conf. Learn. Represent., 2022, pp. 1–11.

[86] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” ACM Comput. Surv., vol. 40, no. 2, pp. 1–60, Apr. 2008.

[87] J. Li, Y. Zhang, H. Shan, and J. Zhang, “Gaitcor: Improved spatial–temporal representation for gait recognition with a hybrid convolution-transformer framework,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Jun. 2023, pp. 1–5.

[88] J. Li, J. Gao, Y. Zhang, H. Shan, and J. Zhang, “Motion matters: A novel motion modeling for cross-view gait feature learning,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Jun. 2023, pp. 1–5.

[89] S. Zagoruyko and N. Komodakis, “Learning to compare image patches via convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 4353–4361.

[90] S. Faigenbaum-Golovin and O. Shimshi, “Image quality assessment: Learning to rank image distortion level,” 2022, arXiv:2208.03117.

[91] X. Liu, J. Van De Wejger, and A. D. Bagdanov, “RankQA: Learning from rankings for no-reference image quality assessment,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1040–1049.

[92] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556.

[93] Y. Meng et al., “Spatial uncertainty-aware semi-supervised crowd counting,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 15529–15539.

[94] X. Wang, Y. Zhan, Y. Zhao, T. Yang, and Q. Ruan, “Semi-supervised crowd counting with spatial temporal consistency and pseudo-label filter,” IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 8, pp. 4190–4203, Aug. 2023.

[95] J. Wan and A. Chan, “Modeling noisy annotations for crowd counting,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 3386–3396.

[96] M. Dai, Z. Huang, J. Gao, H. Shan, and J. Zhang, “Cross-head supervision for crowd counting with noisy annotations,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Jun. 2023, pp. 1–5.

Zhizhong Huang (Graduate Student Member, IEEE) received the B.S. degree from Sichuan University, Chengdu, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Computer Science, Fudan University, Shanghai, China. His research interests include machine/deep learning, computer vision, face aging, and image translation.

Yiming Lei (Member, IEEE) received the bachelor’s degree from Jinzhong University, Jinzhong, China, in 2013, and the M.S. degree from Qingdao University, Qingdao, China, in 2017. He is currently pursuing the Ph.D. degree with the School of Computer Science, Fudan University, Shanghai, China. His research interests include machine/deep learning, computer vision, and biomedical image analysis.

Hongming Shan (Senior Member, IEEE) received the Ph.D. degree in machine learning from Fudan University, Shanghai, China, in 2017. From 2017 to 2020, he was a Postdoctoral Research Associate and a Research Scientist with Rensselaer Polytechnic Institute, Troy, NY, USA. He is currently an Associate Professor with the Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, and a "Qisuao" Research Leader with the Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai. His research focuses on developing machine learning algorithms for biomedical imaging. Dr. Shan was recognized with Youth Outstanding Paper Award at World Artificial Intelligence Conference in 2021.

James Z. Wang (Senior Member, IEEE) received the bachelor’s degree (summa cum laude) in mathematics from the University of Minnesota, Minneapolis, MN, USA, in 1994, and the M.S. degree in mathematics, the M.S. degree in computer science, and the Ph.D. degree in medical information sciences from Stanford University, Stanford, CA, USA, in 1997, 1997, and 2000, respectively.

He was a Visiting Professor with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, from 2007 to 2008. He is a Distinguished Professor of the data sciences and artificial intelligence section with the College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, USA. He is also affiliated with the Department of Communication and Media, School of Social Sciences and Humanities, Loughborough University, Loughborough, U.K., from 2023 to 2024. His research interests include image analysis, affective computing, image modeling, image retrieval, and their applications.

Dr. Wang was a Lead Special Section Guest Editor of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE in 2008, a Program Manager with the Office of the Director of the National Science Foundation from 2011 to 2012, and a Special Issue Guest Editor of the IEEE BIT—The Information Theory Magazine in 2022. He was a recipient of a National Science Foundation Career Award (2004) and Amazon Research Awards (2018-2022).

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Fei-Yue Wang (Fellow, IEEE) received the Ph.D. degree in computer and systems engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990.

He joined The University of Arizona, Tucson, AZ, USA, in 1990, where he became a Professor and the Director of the Robotics and Automation Laboratory and the Program in Advanced Research for Complex Systems. In 1999, he founded the Intelligent Control and Systems Engineering Center, Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Chinese Talents Program from the State Planning Council. In 2002, he was appointed as the Director of the Key Laboratory of Complex Systems and Intelligence Science, CAS. In 2011, he became the State Specially Appointed Expert and the Director of the Key Laboratory for Management and Control of Complex Systems. He is also a Senior Professor and the Chair of the International Academic Advisory Committee for the Institute of Systems Engineering, Macau University of Science and Technology, Macau, China, as well as a Professor with various colleges, including the Artificial Intelligence, and Economics and Management, University of Chinese Academy of Sciences, Beijing, China. His current research focuses on methods and applications for parallel intelligence, social computing, and knowledge automation.

Junping Zhang (Senior Member, IEEE) received the B.S. degree in automation from Xiangtan University, Xiangtan, China, in 1992, the M.S. degree in control theory and control engineering from Hunan University, Changsha, China, in 2000, and the Ph.D. degree in intelligent system and pattern recognition from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2003.

He has been a Professor with the School of Computer Science, Fudan University, Shanghai, China, since 2011. He has widely published in highly ranked international journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and leading international conferences, such as the International Conference on Machine Learning (ICML) and European Conference on Computer Vision (ECCV). His current research interests include machine learning, image processing, biometric authentication, and intelligent transportation systems.

Dr. Zhang has been an Associate Editor of IEEE INTELLIGENT SYSTEMS since 2009.