High-Resolution Crowd Density Maps Generation With Multi-Scale Fusion Conditional GAN

SHAONIAN HUANG1,2, HONGJING ZHOU1, YAO LIU1,3, AND RONGYUAN CHEN1,3

1School of Computer and Information Engineering, Hunan University of Technology and Business, Changsha 410205, China
2Key Laboratory of Hunan Province for New Retail Virtual Reality Technology, Hunan University of Technology and Business, Changsha 410205, China
3Mobile E-Business Collaborative Innovation Center of Hunan Province, Hunan University of Technology and Business, Changsha 410205, China

Corresponding authors: Shaonian Huang (snhuang@hutb.edu.cn) and Rongyuan Chen (chry@hutb.edu.cn)

This work was supported in part by the Scientific Research Fund of Hunan Provincial Education Department under Grant 19A270 and Grant 15b127, and in part by the Social Science Foundation of Hunan Province under Grant 18YBA258.

ABSTRACT The major challenges for density maps estimation and accurate counting stem from the large-scale variations, serious occlusions, and perspective distortions. Existing methods generally suffer from the blurred density maps, which are caused by average convolution kernel, and the ineffective estimation across different crowd scenes. In this paper, we propose a multi-scale fusion conditional generative adversarial network (MFC-GAN) that can generate high-resolution and high-quality density maps. The fusion module of MFC-GAN is embedded in a multi-scale generator and discriminator architecture with a novel adversarial loss, which is designed to guide high-resolution density maps generation. In order to address the problem of scale variation, we further propose a bidirectional fusion module. It combines deep global semantic features and shallow local information by leveraging feature maps presented in different layers of the generator. Furthermore, in order to increase the effectiveness of the multi-scale fusion, we design a cross-attention fusion module, which weights the multi-scale fused feature and learns context-aware feature maps for generating high quality density maps. The experiments on four challenging datasets show the effectiveness, feasibility and robustness of the proposed MFC-GAN.

INDEX TERMS Crowd counting, bidirectional fusion, cross-attention fusion, conditional GAN.

I. INTRODUCTION

With the rapid increasing of urban population and ubiquitous usage of surveillance systems, crowd scene analysis plays an important role in some new applications, such as crowd flow monitoring, assembly controlling, and public safety management [1]. In recent years, the research emphasis on crowd analysis has shifted from simple crowd counting to estimating density maps, which preserve the spatial distribution information of crowd scene [2]. High-quality density maps provide accurate information of crowd scene for various practical applications, such as crowd counting, pedestrian tracking, and crowd stamping prediction. However, due to heavy occlusions, scale and perspective variations, and irregular cluster, it is extremely difficult to generate accurate density value for each pixel of the input crowd image. Inspired by the successful use of convolution neural network (CNN) in image recognition and image segmentation tasks [3]–[6], researchers have proposed various CNN-based network structures to address the issues of density estimation mentioned above [7]–[11]. These methods tend to utilize CNNs with different size of receptive fields to obtain scale aggregation features for adapting the large variation in crowd density.

Although the CNN-based methods for crowd density estimation have been improved tremendously in the past years, most of them suffer from inherent algorithm drawbacks. First, most existing methods employ CNN-based architecture with different convolution kernel size to achieve high performance for crowd counting and density maps estimation. Due to the resizing operation and patch-based training process, these methods usually produce low-resolution feature maps with only 1/8 or 1/16 of the original images. Compared with the performance in low-resolution density maps, the performance of these methods in generating high-resolution density maps may decrease obviously. We design an experiment to demonstrate the performance comparison by using multi-column CNN (MCNN) in [7]. In Fig. 1, we illustrate the 1/8-size...
maps and the original-size maps learned by the MCNN and evaluate them with ShanghaiTech_B [7]. We can see that the high-resolution density maps generated by MCNN do not perform well. We argue that the high-resolution density maps contain more local details, which contribute to the task of counting, tracking, and other related crowd understanding. Most of existing CNN-based methods only use pixel-wise Euclidean loss to optimize models, which has been shown to result in low-quality output on image generation problem [12], [13]. This is because each pixel in the image is assumed to be independent and the Euclidean distance is minimized by averaging all plausible solutions. For crowd images, the averaging operation leads to blurry and over-smoothed crowd density maps. Thus, we propose a high-resolution density maps generation architecture with conditional generative adversarial network as shown in Fig. 2 Motivation1.

Second, some CNN-based methods attempt to address scale variation problem by utilizing multi-path architectures. The MCNN [7] concatenates deep layer feature maps from multi-scale columns directly to generate density maps. SANet [14] employs high-level feature representations with scale aggregation modules to learn high-quality density maps. However, these deep layer features lack low-level spatial information and local details, which are crucial for high-resolution density maps generation. To address these issues, we fuse low-level shallow layer and high-level deep layer to extract more robust feature representations, as illustrated in Fig. 2 Motivation 2.

Last, the low-level layer features usually encode the spatial location information and incorporate some undesirable background noises. Meanwhile, top-level features are really useful for extracting global semantic information, but they fail to describe the spatial distribution information of the crowd. Direct concatenation of the feature maps from different layers leads to the presence of background noise and the absence of local spatial information. This direct feature fusion has no obvious advantage to solve the problem of scale variation in crowd image. To address this issue, we introduce a cross-attention fusion module in the network. This module is used to extract attention-aware feature representation, which is shown in Fig. 2 Motivation 3.

Based on above observations, we present a novel density maps estimation framework called Multi-scale Fusion Conditional Generative Adversarial Network (MFC-GAN). On one hand, inspired by the recent success of conditional GANs in image synthesis [15], we propose an end-to-end density maps generation network with adversarial training loss to alleviate the problem of blurry maps caused by traditional Euclidean loss. The proposed architecture executes a multi-scale density generation and discrimination, which can guarantee the generation of high-resolution density maps. Particularly, our model contains two kinds of complementary density maps generator. One takes small scale input to extract global semantic features, and the other takes larger scale input to enforce local spatial information. On the other hand, we introduce a bidirectional fusion module and a cross-attention fusion module to enforce high-level semantic feature fused with low-level spatial information. Extensive experimental results on four benchmarks have shown the effectiveness of the proposed innovations. The main contributions of this paper are listed as follows:

1) We propose a novel multi-scale fusion conditional GAN architecture to solve the problem of crowd density maps generation in an end-to-end manner. The proposed MFC-GAN can generate high-resolution and high-quality crowd density maps with arbitrary crowd density and perspective.

2) We design a bidirectional fusion mechanism to fuse feature maps from the different layers of the multi-resolution generator. Feature maps extracted from the bidirectional fusion module capture global context information while preserving local spatial details.

3) We present a cross-attention fusion module for combining feature maps obtained from the bidirectional fusion module. Using the strategy of cross-attention, the proposed MFC-GAN achieves obvious improvements in encod-
ing long-range contextual features and suppressing background noises.

4) We conduct extensive experiments on four challenging datasets to validate the effectiveness of the proposed architecture. MFC-GAN achieves superior results over several recent state-of-the-art GAN-based approaches. Furthermore, an ablation study is conducted to demonstrate the improvements obtained by the multi-resolution generator, the bidirectional fusion module, and the cross-attention fusion module.

The rest of this paper is organized as follows. Section II provides a brief overview of the recent work in crowd counting and density maps estimation. Section III describes the proposed framework to generate high-resolution density maps by multi-scale fusion conditional GAN. Section IV presents the implementation details. Section V conducts extensive experiments to evaluate the proposed method. Section VI concludes this paper.

II. RELATED WORK

Various approaches have been proposed to resolve the problem of crowd counting and density maps estimation. They can be roughly divided into three categories: traditional methods, CNN-based regression methods and GAN-based generative methods.

A. TRADITIONAL METHODS

Early crowd counting methods focus on object-based detection to estimate the number of people [16], [17] [18]. Typically, traditional features, such as scale-invariant feature transform (SIFT), histogram oriented gradients (HOG), and edgelet, are extracted to detect heads or bodies. However, these detection-based methods performed poorly in extremely dense crowd scenes. To overcome the occlusion problem in dense crowd scenes, researchers attempted to use regression-based approach to avoid the direct detection [19], [20]. They first extracted a variety of global features and local features, and then employed different regression models to learn a mapping between low-level features and the crowd count. The most commonly used regression techniques include linear regression, Gaussian process regression, and neural network. Regression-based methods successfully solved the problem of occlusion and cluster, but most of them ignored the critical crowd spatial distribution information as they regressed on the global count. In order to utilize the spatial relationship of the local region to estimate the number of people, some researchers attempted to extract different local patch features, and then formulated the density estimation problem as a minimization of a regularized risk quadratic cost function. Pham et al. [21] adopted random forest regression to learn a non-linear mapping between local patch features and density maps. Wang and Zou [22] also proposed a fast method for density estimation based on subspace learning. The comprehensive reviews of crowd counting and density maps estimation methods can be found in the literatures [1], [23], [24].

B. CNN-BASED METHODS

The CNN-based methods in classification, recognition, and segmentation tasks have prompted many researchers to use them for crowd counting and density maps estimation. These CNN-based counting methods learn a mapping from deep feature space to corresponding crowd counts or density maps space. They have achieved superior counting performance compared with traditional methods based on hand-crafted features [1], [25].

Wang et al. [26] first introduced CNN to the task of crowd density maps estimation by using AlexNet architecture [27]. The patch-based deep feature learning was performed in the architecture. Similarly, Shang et al. [28] proposed an end-to-end count estimation framework with the pre-trained GoogLeNet [4] model to extract high-dimensional CNN feature maps. The whole image was taken as an input to generate the final count directly. Recently, different multi-scale architectures were designed to address the issue of scale variation in crowd counting. Zhang et al. [7] designed the MCNN to obtain deep features at various scales. Motivated by MCNN, Sam et al. [9] proposed a switch classifier to select the optimal regressor from multiple independent regressors for input patches. Exploiting dilated convolutional layers to improve scale diversity of features, Li et al. [8] designed a deep network based on VGG-16 [29] to aggregate multi-scale contextual information for density maps estimation. Following the same idea, Cao et al. [14] proposed an encode-decoder network with scale aggregation modules to extract multi-scale feature representations for high-quality density maps generation. To address the accuracy degradation problem of highly crowded scenes, Liu et al. [30] developed an attention-injection deformable convolutional network (ADCrowdNet). Based on the crowd congestion priors detected from an attention-aware network, they conducted a multi-scale estimator to generate high-quality density maps. Liu et al. [31] proposed a deep structured scale integration network (DSSINet) for crowd counting, which used structured feature representation learning and structured loss function optimization to address the scale variation problem. Wan and Chan [32] first studied the impact of different density maps, and then proposed an adaptive density map generator which took the annotation dot map as an input. Liu et al. [33] introduced an end-to-end context-aware crowd counting architecture to tackle the rapid scale change problem. They encoded the scale of the contextual information to predict the crowd density. Lian et al. [34] presented a regression guided detection network (RDNet) for RGB-D crowd counting and localization, which used a depth-adaptive kernel and a depth-aware anchor to facilitate density maps generation in regression and anchor initialization in detection. To fullfill the wide-area crowd counting task, Zhang and Chan [35] employed a multi-view deep neural framework to fuse information from multiple camera views. Sindagi and Patel [36] first started to take high-quality density maps into account. They proposed a contextual pyramid of CNNs (CP-CNN) for generating high-quality
density maps by incorporating global and local contextual information.

C. GAN-BASED METHODS

Inspired by the success of generative adversarial networks in image-to-image translation, some methods introduced the adversarial training model to solve the problem of density maps ambiguity caused by $L_2$ loss. Yang et al. [37] adopted conditional GAN with a multi-scale generator to generate high quality crowd density maps. Shen et al. [38] designed a patch-based crowd counting network with an adversarial training loss and a scale-consistency regularizer to enhance cross-scale density estimation (ACSCP). Li et al. [39] proposed an adversarial learning approach for object counting, which performed adversarial training with pyramid patches of multi-scales from both source-domain and target-domain. To alleviate the burden of labelling data, Olmschenk et al. [40] first introduced a semi-supervised GANs to train crowd counting networks by using minimal training data. Thereafter, Olmschenk et al. [41] proposed a dual-goal GAN for dense crowd counting and real/fake image classification. Recently, Zhou et al. [42] presented a multi-scale generative adversarial network (MS-GAN) for generating high-quality crowd density maps of complex crowd scenes. In the generator of this GAN, a multi-scale fully convolutional network was employed for combining both global and local features. The multi-scale features are learned from different inception modules with different receptive fields to tackle the scale variation problem. The discriminator of this GAN used adversarial loss in additional to Euclidean loss for refining the generated density map. The architecture of this generator is significantly different from the multi-resolution generator of our approach. Most notably, the generator in [42] concatenated the feature maps directly to learn the fusion features from different hierarchy convolutional layers. Our approach adopts bidirectional fusion and cross-attention fusion to aggregate global structure features and local detail features. In addition, we utilize four loss functions for improving the estimation accuracy.

III. MULTI-SCALE FUSION CONDITIONAL GAN

In this section, we cast the problem of density maps estimation as a task of image-to-image translation. The original crowd images and the density maps are regarded as two different image styles. An overview of MFC-GAN for high resolution density maps is given in Fig. 3. It consists of four modules, namely multi-resolution generator, bidirectional fusion module, cross-attention fusion module and multi-scale discriminator. The multi-resolution generator is a CNN-based network for generating coarse-to-fine feature maps. The bidirectional fusion module combines the global semantic feature from deep layers and the local spatial information from shallow layers. The cross-attention fusion module obtains long-distance contextual features with pixel-based attention. The multi-scale discriminator guides generator to produce high-resolution and high quality density maps.

A. MULTI-RESOLUTION GENERATOR

Inspired by the pix2pixHD model in [15], we present a multi-scale conditional GAN architecture to generate high-resolution density maps. The multi-resolution generator consists of four sub-generators $G_\phi$, $G_1^\phi$, $G_2^\phi$, and $G_3^\phi$. It is used to generate coarse-to-fine feature maps as shown in Fig. 3. The global sub-generator $G_\phi$ is used to obtain crowd global context on a coarse scale. It consists of the first ten convolutional layers of VGG-16 [29]. The original image down-sampled by...
the factor of \(8\times\) is passed through the convolutional layers sequentially of \(G_\ell\) to output the coarse feature maps with \(1/8\) of the original resolution.

There are three local sub-generators \(G_1^l, G_2^l,\) and \(G_3^l\), which are used to extract local spatial details at different scales. All the local sub-generators include three components, namely a convolutional front-end, three residual blocks, and a transposed back-end. For the number of residual blocks, we make a tradeoff between accuracy and computing resources. The inputs of the three local sub-generators are down-sampled original image by the factors of \(1\times, 2\times, 4\times\). To integrate the feature maps at different scales, \(G_\ell\) is firstly embedded in \(G_1^l\). The output of \(G_\ell\) summed with the convolutional front-end of \(G_1^l\) is input to the residual blocks of \(G_1^l\). \(G_1^l\) is then integrated into \(G_2^l\) in the same way. Finally, \(G_1^l\) is similarly integrated into \(G_3^l\). We denote the output of the multi-resolution generator as \(C = \{C_0, C_1, C_2, C_3\}\) for the outputs of \(G_\ell, G_1^l, G_2^l,\) and \(G_3^l\). These outputs form multi-resolution pyramid feature maps. The multi-resolution generator structure has been proven to be successful in the image-to-image translation [12], [15]. In this paper, the global sub-generator encodes the global semantic information, and the three local sub-generators capture the local spatial information, and the integration of global and local sub-generators produces coarse-to-fine feature maps.

### B. MULTI-SCALE DISCRIMINATOR

The proposed GAN architecture contains a multi-scale discriminator, which includes three discriminators named \(D_1, D_2,\) and \(D_3\). All the three discriminators share the identical network architecture, and they are trained to differentiate the generated density maps at different scales. Specifically, the generated high-resolution maps are down-sampled by the factors of \(2\times, 4\times, 8\times\) to form a pyramid of three scales. The feature maps of pyramid are first input into \(D_1, D_2,\) and \(D_3\), respectively. Then the discriminators \(D_1, D_2,\) and \(D_3\) are trained to differentiate the ground truth and the generated density maps at different scales. The multi-scale discriminator could guide the multi-resolution generator from coarse-to-fine scales. The top-layer discriminator \(D_3\) with the largest receptive field could guide the global sub-generator from the global view, while the lower-layer discriminators could guide the local sub-generators to concern more about details of image at finer scales.

### C. BIDIRECTIONAL FUSION MODULE

Even through the multi-resolution architecture extracts multi-scale feature maps, the results still lack the details of image structure. One possible reason is that existing generators tend to create a visually realistic image. In other words, generators should focus on the spatial structural information rather than the appearance details. In order to avoid the scale variation problem, generating a high-quality density maps needs to fuse detailed features at different scales. A bidirectional fusion module is implemented to solve the problem of scale variation. It combines the information from multiple layers of the multi-resolution generator.

The proposed bidirectional fusion module is designed to fuse the global context information and the local spatial information at different scales. The whole procedure of the bidirectional fusion module is illustrated in Fig. 4. This module contains two separate fusion paths: the top-down fusion path and the bottom-up fusion path. The top-down path propagates global context information to low-layer features. This path contains three fusions of different scales. In the first scale fusion, the coarse-resolution pyramid feature map \(C_3\) is first upsampled by the factor of 2. Then the upsampled map combines with the pyramid feature map \(C_2\) by element-wise addition to produce a finer feature map \(F_2\). To reduce the channel number of feature maps at different scales, \(C_3\) and \(C_2\) first pass through a \(1 \times 1\) convolutional layer. This process is repeated in the other fusion of two scales until the finest feature map \(F_0\) is obtained. The bottom-up path propagates local spatial information to the high-layer features. Similar to the top-down path, the finest-resolution pyramid feature map \(C_0\) is first downsampled by the factor of 2, and then the downsampled maps add to pyramid feature map \(C_1\) to produce enriched feature map \(F_1\). Finally, the coarse feature map \(F_3\), which integrates with rich location spatial details, is obtained.

### D. CROSS-ATTENTION FUSION MODULE

In the bidirectional fusion module, features produced by the top-down path may contain some background noise. Similarly, features produced by the bottom-up path may excessively suppress the spatial details. Inspired by the successful of residual attention network [43], we consider how to utilize pixel-level attention to combine global coarse feature and local finer feature. We want to extract local spatial details from the top-down fused features to avoid the effectiveness of the background noise. Meanwhile, the global structure information from bottom-up fused features needs to weight local finer feature to obtain density details. Therefore, we design a cross-attention fusion module to combine feature maps from
two paths, which can model long-range contextual features and capture the fine details of crowd image.

Given the set of feature maps \( (F^0_{id}, F^3_{bu}) \) obtained from the bidirectional fusion module, the cross-attention module first concatenates them, and then forwards them to a set of convolutional layers, and finally a sigmoid layer is used to produce attention maps. The attention maps can be described as

\[
F_A = \text{Sigmoid}(\text{cat}(F^0_{id}, F^3_{bu})).
\]

The obtained attention maps are then weighted to the bidirectional fused feature maps via element-wise multiply operation. The result is summed with the bidirectional fused features and the cross-attention based feature maps are given by

\[
F_f = (1 + F_A) \odot F^0_{id} + (1 + F_A) \odot F^3_{bu}.
\]

where \( \odot \) denotes element-wise multiplication. Fig. 5 shows the cross-attention fusion module. The final feature maps \( F_f \) are operated with \( 1 \times 1 \) convolutional layer to produce the final density map.

E. LOSS FUNCTION

In order to optimize the proposed architecture, we utilize four loss functions, namely the adversarial loss \( L_A(G, D) \), the feature matching loss \( L_{FM}(G, D) \), the local pattern consistency loss \( L_{LC}(G) \) and the Euclidean loss \( L_E(G) \). We use the enhanced adversarial loss with feature matching loss to stabilize the training process. These two losses force the multi-resolution generator to capture the global image structure from the multi-scale feature maps. The local pattern consistency loss and the Euclidean loss are used to reinforce the fine features and preserve the detail information. The final objective function combines the above four loss functions as

\[
L = L_A(G, D) + \lambda_{fm} L_{FM}(G, D) + \lambda_{e} L_E(G) + \lambda_{lc} L_{LC}(G),
\]

where the hyperparameters \( \lambda_{fm} \), \( \lambda_{e} \), and \( \lambda_{lc} \) are predefined weights for feature matching loss, Euclidean loss, and local pattern consistency loss, respectively.

1) ADVERSARIAL LOSS

We adopt the adversarial loss of conditional GAN. It consists of a multi-resolution generator \( G \) and a multi-scale discriminator \( D \). The objective of the generator \( G \) in our task is to translate a crowd image to a high-resolution density map. The discriminator \( D \) aims to distinguish whether a density map is real or fake. To differentiate the original high-resolution image and the generated density map, we adopt a multi-scale discriminator \( D = \{D_1, D_2, D_3\} \), which forces the generator to produce coarse-to-fine features. With the multi-scale discriminator, the adversarial loss becomes a multi-task learning problem of

\[
L_A(G, D) = \min_G \max_{D_1, D_2, D_3} \sum_{k=1,2,3} L_{GAN}(G, D_k)
\]

where \( L_{GAN}(G, D_k) \) is the adversarial loss of the \( k \)-th discriminator. Its objective function is

\[
L_{GAN}(G, D_k) = E_{(x,y)}[\log D_k(x, y)] + E_{(x)}[\log(1 - D_k(x, G(x))]
\]

where \( x \) and \( y \) denote the original crowd image and the corresponding density map, \( G(x) \) represents the output produced by the generator of the GAN network, \( E_{(x,y)} \) denotes the expectation of the joint distribution of \( x \) and \( y \), and \( E_{(x)} \) denotes the expectation of \( x \).

2) FEATURE MATCHING LOSS

Feature matching loss and perceptual loss [12], [15] [44] have been verified to be useful for image super-resolution and image synthesis. In order to save computing resource, we incorporate feature matching loss based on the discriminator to improve the GAN loss. This loss guides the generator to produce nature image base on multi-scale statistical information. The feature matching loss is formulated as

\[
L_{FM}(G, D_k) = \sum_{k=1,2,3} \ell_{FM}(G, D_k)
\]

where \( \ell_{FM}(G, D_k) \) is the feature matching loss of the discriminator \( D_k \). Specifically, the intermediate feature maps from multiple layers of the discriminator \( D_k \) are learned to match between the crowd image and the generated density map. The feature matching loss \( \ell_{FM}(G, D_k) \) is then calculated as

\[
\ell_{FM}(G, D_k) = E_{(x,y)} \sum_{i=1}^{L} \frac{1}{N_i} \| D_k^i(x, y) - D_k^i(x, G(x)) \|_1
\]

where \( L \) is the total number of layers for intermediate fea-

3) EUCLIDEAN LOSS

The Euclidean loss is chosen to force the generated density map to be close to the ground truth. It is defined as

\[
L_E(G) = \frac{1}{N} \| G(x) - y \|^2_2.
\]
where \( N \) is the number of pixels in density map, \( x \) is the input crowd image, \( y \) is the corresponding ground truth, and \( G(x) \) denotes the estimated density map. The Euclidean loss measures estimation error in a pixel-level manner.

4) LOCAL PATTERN CONSISTENCY LOSS
In order to improve the quality of density maps, we incorporate multi-scale Structure Similarity Index (MS_SSIM) [45] to measure the local consistency between estimated density maps and ground truths. SSIM is an indicator which is wildly used to compute the similarity between two images in terms of local luminance, contrast, and structure comparison. SANet [14] and CCWild [46] adopt SSIM to generate high-quality density maps for crowd counting. MS_SSIM is an extension of SSIM. The overall MS_SSIM evaluation is calculated by combining crowd image similarity at various scales using

\[
MS_{SSIM}(x, y) = [\ell_m(x, y)]^{a_m} \sum_{i=1}^{m} [c_i(x, y)]^{b_i} [s_i(x, y)]^{y_i},
\]

where \( \ell_m(x, y) \) denotes the luminance comparison which is computed only at one scale, \( c_i(x, y) \) and \( s_i(x, y) \) represent the contrast comparison and the structure comparison at the \( i \)-th scale, respectively. The values of exponents \( a_m, b_i, y_i \) are the same as those in [45]. The value of MS_SSIM ranges from 0 to 1. MS_SSIM equals to 1 when the two images are identical. Finally, the local pattern consistency loss is given by

\[
L_{LC}(G) = 1 - \frac{1}{N} \sum_x MS_{SSIM}(G(x), y),
\]

where \( N \) is the number of pixels in the density map, \( G(x) \) is the generated density map and \( y \) is the corresponding ground truth.

IV. IMPLEMENTATION DETAILS
In this section, we describe the specific implementation details of MFC-GAN.

A. TRAINING DETAILS
In the training stage, the crowd images of original size are horizontally flipped randomly for data augmentation. For training the network, we first need to convert the human head annotations into corresponding density maps. Following the method of generating density maps in [7], we adopt geometry-adaptive kernels to tackle the images of crowd scene. The ground truth is defined as

\[
y(x) = \sum_{i=1}^{N} \delta(x - x_i) \times G_{\sigma_i}(x),
\]

where \( \sigma_i = \beta \overline{d_i} \) and \( \overline{d_i} = \frac{1}{k} \sum_{j=1}^{k} d_j^i \). For each target head \( x_i \) in the input, \( \overline{d_i} \) indicates the average distance to its \( k \) nearest neighbors. In the experiments, we follow the setting in [7] where \( \beta = 0.3 \) and \( k = 3 \). We then train the MFC-GAN in an end-to-end manner. The global generator weights are fine-tuned from a well-trained VGG-16 [29]. Other parameters are randomly initialized by a Gaussian distribution. The mean is 0 and the standard deviation is 0.01. Adam optimizer [47] with learning rate of \( 2e^{0.4} \) and a momentum of 0.9 is used to train the model. The exponential decay rate \( \beta_1 \) is set to 0.5 and \( \beta_2 \) is set to 0.999. The hyper-parameter \( \lambda_{fn} \) and \( \lambda_{sk} \) in Equation(3) is set to 10. The hyper-parameter \( \lambda_{e} \) in Equation(3) is set to 500. Each batch contains only one randomly selected image to reduce the computational burdens of training model. All the experimental trainings and evaluations are implemented with the PyTorch platform [48].

B. EVALUATION METRICS
We evaluate MFC-GAN architecture in two aspects: the counting performance and the quality of density maps. For counting performance, Mean Absolute Error (MAE) and Mean Squared Error (MSE) are commonly used metrics in previous methods. They can be computed as:

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |Y_i - Y_i^{GT}|,
\]

\[
MSE = \frac{1}{N} \sum_{i=1}^{N} \|Y_i - Y_i^{GT}\|^2,
\]

where \( N \) is the number of images in one test sequence. \( Y_i \) and \( Y_i^{GT} \) are the estimated counting numbers and the ground truth counting numbers of the \( i \)-th test image. \( Y_i \) is given by the integration of the estimated density map. It can be calculated by

\[
Y_i = \sum_{i=1}^{w} \sum_{j=1}^{h} y(i, j),
\]

where \( w \) and \( h \) are the width and height of the generated density map respectively, \( y(i, j) \) denotes the pixel value at \( (i, j) \) of the generated density map.

In order to evaluate the quality of generated density maps, we also use the PSNR and SSIM. The calculating method of them is the same as that in [36], [45].

V. EXPERIMENTS
In this section, we demonstrate the effectiveness of the proposed MFC-GAN on four popular crowd counting datasets. We first introduce the four datasets, and then an ablation study is reported to demonstrate the improvements of the different modules involved in the proposed architecture. It is followed by a detailed evaluation results and performance comparisons with recent state-of-the-art methods.

A. DATASETS
We evaluate our MFC-GAN architecture on four crowd counting datasets: ShanghaiTech [7], UCF_CC_50 [49], UCF_QNRF [50], UCSD [51].
TABLE 1. Ablation study settings.

| Method      | Global | Local | TDF  | BDF | CAF |
|-------------|--------|-------|------|-----|-----|
| GAN1        | ✓      | ✓     | ✓    | ✓   | ✓   |
| GAN2        | ✓      | ✓     | ✓    | ✓   | ✓   |
| GAN2+TDF    | ✓      | ✓     | ✓    | ✓   | ✓   |
| GAN2+BDF    | ✓      | ✓     | ✓    | ✓   | ✓   |
| MFC-GAN     | ✓      | ✓     | ✓    | ✓   | ✓   |

1) SHANGHAITECH
The ShanghaiTech crowd counting dataset contains 1198 annotated images with a total number of 330,165 persons [7]. It is divided into two parts as Part_A with 482 images and Part_B with 716 images. Part_A contains highly congested crowd scenes randomly collected from the Internet. Part_B contains relatively sparse crowd images captured from crowd streets in Shanghai. The two parts are further split into training datasets and testing datasets. Part_A has 300 images for training and 182 images for testing. Part_B has 400 images for training and 316 images for testing. For both subsets, ground truth density maps are generated with geometry-adaptive kernels to alleviate perspective distortion.

2) UCF_CC_50
UCF_CC_50 [49] contains 50 extremely dense crowd images crawled from the Internet. The number of annotated individuals per image ranges from 94 to 4553 with an average number of 1280. Counting people in this dataset is challenging due to the large variability of resolution, perspective, and density. In the experiments, we use 5-fold cross-validation followed the standard setting in [49] to evaluate the performance of our method. We generate ground truth density maps with geometry-adaptive kernels.

3) UCF_QNRF
UCF_QNRF is a larger-scale crowd dataset introduced by Idrees et al. [50]. This dataset contains 1,535 images with 1.25 million annotations collected from web sites. This dataset is dense density and high-resolution with average counting number of 815 and average resolution of 2013 × 2902. It has diverse viewpoint, lighting variation, and backgrounds. The training set and testing set consist of 1201 and 334 images, respectively. We follow the experiments settings of [36] to generated density maps.

4) WORLDEXPO’10
The WorldExpo’10 dataset contains 3980 annotated video frames captured by 108 surveillance cameras from Shanghai 2010 WorldExpo [52]. These video frames selected from five different scenes are divided into a training set with 3380 frames and a testing set with 600 frames. The region of interest (ROI) and perspective maps are provided for the dataset. Follow the same experimental setting in [36], only the ROI regions are taken into consideration during preprocessing, and the perspective maps are used to generate the ground truth.

B. ABLATION STUDY
To better understand the effectiveness of the architecture of our method, we perform a detailed ablation study by considering the combination of four factors: global generator, local generators, bidirectional fusion module, and cross-attention fusion module. We conduct the following variants with different components combination: 1) GAN1: only a global generator is used in the GAN architecture; 2) GAN2: a global generator and three local generators are used; 3) GAN2+TDF: the dop-down path fusion module followed the GAN2 network; 4) GAN2+BDF: the bidirectional fusion module followed the GAN2 network; 5) MFC-GAN: the cross attention fusion (CAF) module followed the GAN2+BDF network. The ablation configurations are listed in Table 1. We compare the performance of our proposed method with that of GAN1, GAN2, GAN2+TDF and GAN2+BDF on the dataset of ShanghaiTech_A and ShanghaiTech_B. The quantitative and qualitative results of the ablation study are shown in Table 2 and Fig. 6. They demonstrate that the proposed architecture achieves the best performance of density maps in both MAE and MSE. In addition, to validate the contribution of each loss in the Equation (3), we also conduct an ablation study by using different combinations of losses. The results of the ablation study for loss function are presented in Table 3.

1) ABLATION FOR MULTI-RESOLUTION GENERATION MECHANISM
We separately investigate the role of multi-resolution generation mechanism in the proposed method. First, we train the network without multi-resolution generation mechanism. There only exists a global generator to generate density maps. Then, we combine a global generator and three local generators to obtain high-resolution density maps. The comparison results are shown in Table 2. It can be observed that the GAN2 improves the performance by almost 23.4% in MAE on ShanghaiTech_A and 35.3% in MAE on ShanghaiTech_B compared with the results of GAN1.

2) ABLATION FOR BIDIRECTIONAL BASED FUSION STEPSWISE
We construct a GAN2+TDF network and a GAN2+BDF network to study the performance of the proposed bidirectional fusion module on on ShanghaiTech_A and ShanghaiTech_B datasets. Table 2 reports the comparison results.
FIGURE 6. Ablation study results: (a) Input. (b) GAN1. (C) GAN2. (d) GAN2+TDF. (e) GAN2+BDF. (f) MFC-GAN. (g) Ground truth density map.
of GAN+TDF, GAN+BDF and GAN2. It can be observed that GAN2+BDF architecture outperforms GAN2 by nearly 26% in MAE on Shanghaitech_A and 29.7% in MAE on Shanghaitech_B. We also observe that GAN2+BDF network improves over GAN2+TDF network by 2% in MAE on Shanghaitech_A and 8% in MAE on Shanghaitech_B.

3) ABSTRACTION FOR CROSS-ATTENTION BASED FUSION STEPWISE

We also train the MFC-GAN network to study the performance of the proposed directional fusion module and the cross-attention fusion module. As shown in Table 2, the proposed MFC-GAN architecture improves over GAN2+BDF network by 13.7% in MAE on Shanghaitech_A and 32.3% in MAE on Shanghaitech_B.

Fig. 6 shows the qualitative results of the ablation study. It can be observed from Fig. 6(b) and Fig. 6(c) that the direct utilization of deep feature maps results in a lot of background noise and a low accuracy in final crowd counting. The GAN2+BDF approach shown in Fig. 6(e) results in a higher counting accuracy and the refined density maps. However, they still loss some details in the final density maps. Fig. 6(f) shows that the proposed MFC-GAN architecture achieves the closest results to the ground truth with much lesser noise clutters and the best performance of density maps in PSNR and SSIM. It can be observed from Table 2 and Fig. 6 that the bidirectional fusion module and the cross-attention fusion module can greatly boost the final counting performance by using the proposed multi-resolution generation mechanism.

4) ABSTRACTION FOR LOSS FUNCTION

We run a ablation study on the choice of different losses in Equation (3) for training the generator. The results are shown in Table 3. From the first three rows of Table 3, we can see that the combination of adversarial loss and Euclidean loss significantly improves the counting performance compared to that using only adversarial loss or Euclidean loss. As shown in the 3rd and 4th rows of Table 3, the MAE performance increases by nearly 3% when the feature matching loss is incorporated. This validates the claim that the feature matching loss promotes the generator to produce finer image from multi-scale feature maps. As shown in the 3rd and 5th rows of Table 3, the MAE performance increases by 11.3% by using the local pattern consistency loss to train the network. The increased MAE demonstrates the necessity of the local pattern consistency loss in the generator loss function. Based on the above results of experiments, the network rationality of the proposed method has been verified.

C. COMPARISONS WITH STATE-OF-THE-ART

We demonstrate the efficiency of MFC-GAN with other existing methods on four challenging crowd counting datasets, namely Shanghaitech, UCF_CC_50, UCF_QNRF and UCSD datasets.

TABLE 3. Ablation study on Shanghaitech_B for the density map generator using different loss functions. “Adv”, “Euc”, “Feat”, and “Lc” denote adversarial loss, Euclidean loss, feature matching loss, and local pattern consistency loss in Equation (3), respectively.

| Loss | MAE | MSE |
|------|-----|-----|
| Adv  | 63.8 | 103.8 |
| Euc  | 25.5 | 41.2 |
| Euc+Adv | 14.2 | 23.4 |
| Euc+AdvFeat | 13.8 | 22.5 |
| Euc+Adv+Lc | 12.6 | 20.7 |
| Euc+Adv+Feat+Lc | 10.9 | 17.5 |

TABLE 4. Comparison results on Shanghaitech dataset [7]. The best performance and the results of the proposed method are highlighted in blue and bold respectively.

| method        | Shanghaitech_A | Shanghaitech_B |
|---------------|----------------|----------------|
|               | MAE | MSE | MAE | MSE |
| Zhan et al. [52] | 181.8 | 277.7 | 32.0 | 49.8 |
| MCNN [7]       | 110.2 | 173.2 | 26.4 | 41.3 |
| CP-CNN [36]    | 73.6 | 106.4 | 20.1 | 30.1 |
| MS-GAN [42]    | 96.5 | 135.9 | 18.7 | 30.5 |
| ACSCP [38]     | 75.7 | 102.7 | 17.2 | 27.4 |
| DANet [10]     | 71.6 | 104.9 | 15.0 | 21.9 |
| ADCrowdNet [30] | 70.9 | 115.2 | 7.7 | 12.9 |
| Wan et al. [32] | 64.7 | 97.1 | 8.1 | 13.6 |
| DSSI-Net [31]  | 60.6 | 96.0 | 6.9 | 10.3 |
| MFC-GAN        | **70.1** | **107.2** | **10.9** | **17.5** |

TABLE 5. Quality of density maps on ShanghaiTech_A dataset.

| Method  | PSNR | SSIM |
|---------|------|------|
| MCNN [7] | 21.40 | 0.52 |
| ACSCP [38] | 22.30 | 0.67 |
| CP-CNN [36] | 21.72 | 0.72 |
| CSRNet [8] | **23.79** | 0.76 |
| MFC-GAN | **23.63** | **0.77** |

1) SHANGHAITECH [7]

We compare our method with several typical crowd counting methods on Shanghaitech_A and Shanghaitech_B datasets: Zhang et al. [52], MCNN [7], CP-CNN [36], MS-GAN [42], ACSCP [38], DANet [10], ADCrowdNet [30], Wan and Li [32], DSSI-Net [31]. These methods include the CNN-based regression methods and the GAN-based generative methods. As shown in Table 4, our method achieves better performance on both parts of the Shanghaitech dataset compared to the two GAN-based methods ACSCP [38] and MS-GAN [42]. Specifically, on Shanghaitech_B, our method achieves an improvement of 36.6% in MAE and 36.1% in MSE over the existing best GAN-based method ACSCP. Compared with the CNN-based regression methods, the results of MFC-GAN outperform four methods in the Table 4. We also note that three state-of-the-art methods have achieved excellent performance on this dataset, such as DSSI-Net [31], Wan and Li [32] and ADCrowdNet [30]. All these three methods are CNN-based regression methods and adopt patch-based training process.

To evaluate the quality of generated density maps, we compare our method to the MCNN [7], CP-CNN [36],
CSRNet [8] and ACSCP [38] using Shanghaitech_A dataset. We adopt PSNR and SSIM as evaluation metrics. Figure 7 illustrates the qualitative comparison results of the CSRNet, ACSCP and the proposed MFC-GAN. The density maps generated by ACSCP method contain some obvious noise, and the estimated counts are quite biased. The quantitative evaluation results are shown in Table 5. We can see from Table 5 that MFC-GAN achieves the highest SSIM.

2) UCF_CC_50 [49]

In Table 6, the proposed method is evaluated against seven crowd counting methods: Zhang et al. [52], MCNN [7], MS-GAN [42], CP-CNN [36], ACSCP [38], DANet [10], ADCrowdNet [30], DSSINet [31] in MAE and MSE metrics. As shown in Table 6, the proposed method gets the lowest MSE and a comparable MAE compared to seven state-of-the-art methods on UCF_CC_50 dataset. Specifically, our method obtains an improvement of 17.9% in MAE and 25.9% in MSE over the existing best GAN-based method ACSCP. Meanwhile, our method achieves an improvement of 30.9% in MAE and 28.3% in MSE compared to the GAN-based method MS-GAN. Qualitative results for sample images from UCF_CC_50 dataset are presented in Fig. 8. It can be seen that there is a certain deviation between the estimated counting and the ground-truth.

3) UCF_QNRF [50]

The proposed method is evaluated against six recent CNN-based approaches: MCNN [7], CMTL [53], Idrees et al. [50], HA-CCN [54], Wan and Li [32], DSSINet [31]. We can see that the results of our method

FIGURE 7. Comparison of the estimated density maps. (a) Test images. (b) Estimated density maps by CSRNet [8]. (c) Estimated density maps by ACSCP [38]. (d) Estimated density maps by the proposed MFC-GAN. (e) Ground-truth density maps with crowd count.
outperform two CNN-based methods in Table 7. It states that the proposed GAN-based method has an acceptable performance in the case of extremely dense large-scale dataset. Some example results of the proposed method are shown in Fig. 9. Three images with different crowd density are chosen from UCF_QNRF testing set. We find that the proposed method has achieved acceptable results on complex crowd scenes with significantly density variation.

4) WorldExpo’10 [52]
We compare the proposed method with eight typical crowd counting methods on WorldExpo dataset: Zhang et al. [52], MCNN [7], Shang et al. [28], Wan and Li [32], CP-CNN [36], DANet [10], ADCrowdNet [30], DSSINet [31]. Results are shown in Table 8. It can be observed that the proposed MFC-GAN remains comparable performance in terms of

### FIGURE 8. Qualitative results of the proposed method on UCF_CC_50.

![Input](image1)

![Ground truth](image2)

![Estimation](image3)

(a) Input  
(b) Ground truth  
(c) Estimation

### FIGURE 9. Visualization of the crowd density maps generated by the proposed method on UCF_QNRF.

![Image1](image4)

![Image2](image5)

![Image3](image6)

(a)  
(b)  
(c)

(a) Input  
(b) Ground truth  
(c) Estimation

### TABLE 6. Comparison results on UCF_CC_50 dataset [49].

| Method         | MAE  | MSE  |
|----------------|------|------|
| Zhan et al.    | 467.0| 498.5|
| MCNN [7]       | 377.6| 509.1|
| MS-GAN [42]    | 345.7| 418.3|
| CP-CNN [36]    | 298.5| 320.9|
| ACS [38]       | 291.0| 404.6|
| DANet [10]     | 290.8| 326.5|
| ADCrowdNet [30]| 273.6| 362.0|
| DSSINet [31]   | 216.9| 302.4|
| MFC-GAN        | 239.0| 299.9|
We notice that the proposed MFC-GAN outperforms the two existing GAN-based methods in the quantitative evaluation for four public crowd counting datasets. However,
several state-of-art CNN-based methods achieve better multi-scale performance on different datasets, such as ADCrowdNet [30], DSSINet [31] and Wan and Li [32]. From Table 4, we can see that these three methods have achieved excellent performance on ShanghaiTech dataset. Table 7 shows that DSSINet [31] and Wan and Li [32] achieve better performance in MAE and MSE on UCF_QNRF dataset. As shown in Table 8, on the WorldExpo’10 dataset, the average MAE of DSSINet [31] and ADCrowdNet [30] come to the top and the second place. The main reasons for these phenomena are: 1) DSSINet [31] adopts three subnetworks to train the overall framework. Each subnetwork is composed of the first ten convolutional layer of VGG16. Thus, the depth of the DSSINet is deeper than the generator of proposed MFC-GAN. This helps it to refine the multiscale features of crowd scene. 2) ADCrowdNet [30] and Wan and Li [32] both use different types of prior knowledge to train the network for crowd counting. ADCrowdNet [30] pre-trains attention map generator to obtain the prior of candidate crowd regions and the prior of congestion degree in crowd regions. Wan and Li [32] uses the generated density map as prior knowledge to refine the estimated density map. The generated density maps are learned by a self-attention fusion network. But the proposed MFC-GAN achieves comparable performance without any prior knowledge. Despite its low performance, the advantages of the proposed method over the above methods are two-fold: 1) MFC-GAN adopts an end-to-end manner to generate high-resolution density map. The above CNN-based methods use patch-based manner to train the model, and then predict each sliding window during the testing phase. The counting number of different sliding windows need to be assembled to obtain the final total count of the image. This mechanism may lead to deviations in counting results. The proposed MFC-GAN takes the whole image as input, and outputs corresponding density map and a total number of the crowds. 2) The proposed MFC-GAN proves the effectiveness of GAN-based method in the crowd counting task. By combining different loss functions, the discriminator of MFC-GAN forces the generator to gradually produce high-quality density map closed to the ground-truth. At the same time, the architecture of the MFC-GAN generator can be easily adjusted to different scales to adapt different crowd scenes.

We also discuss the complexity of our method. Table 9 shows the amount of parameters of four crowd counting methods including the proposed MFC-GAN. As listed in Table 9, the number of parameters of MFC-GAN is 29.6 million. MFC-GAN has more parameters than the GAN-based method ACSCP, but MFC-GAN achieves better performance on four crowds counting datasets. Compared with the CP-CNN method focusing on high-quality density map generation, MFC-GAN achieves better performance with fewer parameters. During the testing phrase, MFC-GAN takes 3.08 s to produce a 1024 × 768 density map on an NVIDIA 1080 GPU.

### VI. CONCLUSION

In this paper, we proposed an end-to-end architecture called multi-scale fusion conditional generative adversarial network (MFC-GAN) for high-resolution and high-quality crowd density maps. We designed a multi-resolution conditional GAN to generate high-resolution density maps generation. In order to address the scale variation and the spatial details problem, we proposed a bidirectional fusion module and a cross-attention fusion module, which can extract the multi-scale fused feature maps weighted by the pixel-level attention maps. In this way, the proposed architecture could combine global semantic feature and local spatial details to generate high-quality density maps. The experimental results on several public challenging datasets have shown that our approach outperforms many typical density maps estimation methods, and the ablation study demonstrate the effectiveness of the proposed architecture. In the future, we will explore better ways of fusing features of different layers and extent the proposed architecture to the other tasks of crowd understanding. We will also explore how to extract more effective domain-invariant features of different crowd scenes, and further expand MFC-GAN for cross-scene counting.

### REFERENCES

[1] V. A. Sindagi and V. M. Patel, “A survey of recent advances in CNN-based single image crowd counting and density estimation,” *Pattern Recognit. Lett.*, vol. 107, pp. 3–16, May 2018.

[2] C. Zhang, K. Kang, H. Li, X. Wang, R. Xie, and X. Yang, “Data-driven crowd understanding: A baseline for a large-scale crowd dataset,” *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1048–1061, Jun. 2016.

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[5] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeePLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[7] 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.

[8] 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.

[9] 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.

[10] Z. Zou, X. Su, X. Qu, and P. Zhou, “DA-Net: Learning the fine-grained density distribution with deformation aggregation network,” *IEEE Access*, vol. 6, pp. 60745–60756, 2018.

[11] W. Kong, H. Li, G. Xing, and F. Zhao, “An automatic scale-adaptive feature fusion for understanding the highly congested scenes,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1091–1100.

[12] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 694–711.
[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image translation with conditional adversarial networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

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

[15] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional GANs,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

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

[17] L. Wang, L. Xu, and J.-H. Yang, “Pedestrian detection in crowded scenes via scale and occlusion analysis,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1210–1214.

[18] O. Tuzel, F. Porikli, and P. Meer, “Pedestrian detection via classification on Riemannian manifolds,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1713–1727, Oct. 2008.

[19] 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.

[20] K. Chen, C. C. Loy, S. Gong, and T. Xiang, “Feature mining for localised crowd counting,” in *Proc. Proceedings Brit. Mach. Vis. Conf.*, 2012, vol. 1, no. 2, p. 3.

[21] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, “COUNT forest: CO-voting uncertain number of targets using random forest for crowd density estimation,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3253–3261.

[22] Y. Wang and Y. Zou, “Fast visual object counting via example-based density estimation,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3653–3657.

[23] S. A. M. Salem, S. A. Saundi, and H. Ibrahim, “Recent survey on crowd density estimation and counting for visual surveillance,” *Eng. Appl. Artif. Intell.*, vol. 41, pp. 10–114, May 2015.

[24] B.-Z. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu, “Crowd analysis: A survey,” *Mach. Vis. Appl.*, vol. 19, nos. 5–6, pp. 345–357, Oct. 2008.

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

[26] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, “Deep people counting in extremely dense crowds,” in *Proc. 23rd ACM Int. Conf. Multimedia (MM)*, Nov. 2015, pp. 323–335, 2020.

[27] X. Cao, Z. Wang, Y. Zhao, and F. Su, “Scale aggregation network for accurate and efficient crowd counting,” in *Proc. 23rd ACM Int. Conf. Multimedia (MM)*, Nov. 2015, pp. 323–335, 2020.

[28] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3156–3164.

[29] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, “Perceptual generative adversarial networks for small object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1222–1230.

[30] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, pp. 1398–1402.

[31] Q. Wang, J. Gao, W. Lin, and Y. Yuan, “Learning from synthetic data for crowd counting in the wild,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 8198–8207.

[32] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[33] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, “Multi-source multi-scale counting in extremely dense crowds,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2547–2554.

[34] H. Idrees, M. Tazhayakov, M. Chen, D. Zhang, S. Al-Marzooq, N. Rajpoot, and M. Shah, “Composition loss for counting, density map estimation and localization in dense crowds,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 532–546.

[35] A. B. Chan, Z.-S. John Liang, and N. Vasconcelos, “Privacy preserving crowd monitoring: Counting people without people models or tracking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[36] V. A. Sindagi and V. M. Patel, “CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting,” in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6.

[37] V. A. Sindagi and V. M. Patel, “HA-CCNN: Hierarchical attention-based crowd counting network,” *IEEE Trans. Image Process.*, vol. 29, pp. 323–335, 2020.

SHAOHUI HUANG received the B.S. degree in computer science from Hunan Normal University, China, in 2000, and the M.S. and Ph.D. degrees in computer science from Central South University, China, in 2006 and 2018, respectively. She is currently an Associate Professor with the School of Computer and Information Engineering, Hunan University of Technology and Business, China. Her current research interests include computer vision, machine learning, and crowd behavior analysis and modeling.
HONGJING ZHOU received the B.S. degree in computer science from South-Central College for Nationalities, China, in 1998, and the M.S. degree in software engineering from Hunan University, China, in 2009. She is currently a Lecturer with the School of Computer and Information Engineering, Hunan University of Technology and Business, China. Her current research interests include opportunistic networks, mobile social networks, and machine learning.

YAO LIU received the B.S. degree in computer science from the Central South University of Technology, China, in 1998, and the M.S. degree in pattern recognition and intelligence system and the Ph.D. degree in computer science from Central South University, China, in 2004 and 2012, respectively. He is currently an Associate Professor with the School of Computer and Information Engineering, Hunan University of Technology and Business, China. His current research interests include opportunistic networks, mobile social networks, and mobile edge computing.

RONGYUAN CHEN received the B.S. degree in computer science from Nanjing University, China, in 2000, the M.S. degree from the Changsha University of Science and Technology, China, in 2005, and the Ph.D. degree in computer science from Wuhan University, China, in 2010. He is currently a Professor with the School of Computer and Information Engineering, Hunan University of Technology and Business. His current research interests include image processing and machine learning.

* * *

IEEEAccess