Crowd Density Estimation Using Fusion of Multi-Layer Features

Xinghao Ding, *Member, IEEE*, Fujin He, Zhirui Lin, Yu Wang, Huimin Guo, and Yue Huang

***Abstract*— Crowd counting is very important in many tasks such as video surveillance, traffic monitoring, public security, and urban planning, so it is a very important part of the intelligent transportation system. However, achieving an accurate crowd counting and generating a precise density map are still challenging tasks due to the occlusion, perspective distortion, complex backgrounds, and varying scales. In addition, most of the existing methods focus only on the accuracy of crowd counting without considering the correctness of a density distribution; namely, there are many false negatives and false positives in a generated density map. To address this issue, we propose a novel encoder-decoder Convolution Neural Network (CNN) that fuses the feature maps in both encoding and decoding sub-networks to generate a more reasonable density map and estimate the number of people more accurately. Furthermore, we introduce a new evaluation method named the Patch Absolute Error (PAE) which is more appropriate to measure the accuracy of a density map. The extensive experiments on several existing public crowd counting datasets demonstrate that our approach achieves better performance than the current state-of-the-art methods. Lastly, considering the cross-scene crowd counting in practice, we evaluate our model on some cross-scene datasets. The results show our method has a good performance in cross-scene datasets.**

***Index Terms*— Crowd counting, fusion, encoder-decoder, density map.**

I. Introduction

With the development of society, the density of urban population continues to increase, which leads to a large number of the overcrowded situations, such as those in the subway stations, bus stations, airports, shopping malls, tourist attractions or other places intended for the large-scale activities, Effective crowd management is meaningful for intelligent transportation systems [1], [4], [14], [47]. In order to avoid the overcrowding problems, the intensity of crowds needs to be monitored to react timely and disperse the crowd. With the popularity of video surveillance and deployment of vision technology, a great interest has appeared in the crowd scene analysis in various environments. This paper analyzes mainly the crowd counting and high-quality density maps generation, which are important in the applications such as traffic monitoring, public security, urban planning, flow monitoring [38], [39], etc. It should be mentioned that crowd counting is a total different task with semantic segmentation. Semantic segmentation is an accurate pixel-level classification task that requires accurate pixel-level labels, which is different from the crowd counting. Crowd counting is casted as a complicated fuzzy function mapping problem, and then distribution map is one of the widely-used estimators [6], [24], [35]. In the proposed task, we only need to count how many people in a region, without generating the pixel-level contour of the targets.

The taxonomy of the traditional crowd counting algorithms consists of two paradigms: model-based detection and feature-based regression [19]. Model-based detection algorithms detect and count people individually, and they are usually based on the motion features, foreground segmentation, silhouette/shape matching, and object recognition methods [17], [22], [40]. Typically, a pedestrian is considered as an individual entity that can be detected by a sliding-window [12], [27] detector. The low-level handcrafted features (e.g. HOG [9], SIFT [24] and haar-like [23]) are used to train the pedestrians' classifiers. On the other hand, the feature-based regression approaches extract the features such as foreground pixels, interest points, texture and vectors formed with those features, and then learn a regression function to estimate the crowd density or human count [5], [7], [16], [21], [44]. In [7] the authors found that spatio-temporal information is effective for improving the performance of crowd counting. Based on this observation, the author exploits spatiotemporal information into multi-linear regression learning, achieving accurate crowd counting. However, these traditional methods [5], [7], [16], [21], [44] suffer from low performance in complex situations, containing the significant occlusion, non-uniform illumination, perspective effects and large variability of scales.

Recently presented deep learning methods [13], [29], [30], e.g. CNN based methods [11], [43], have brought a large improvement in performance compared with the traditional crowd counting algorithms based on the handcrafted features. Inspired by the success of multi-task learning [41], [45], [50] in various computer vision tasks, Zhang *et al.* [46], and Simonyan and Zisserman [37] combined the crowd counting

Manuscript received July 24, 2018; revised May 7, 2019, October 1, 2019, and December 30, 2019; accepted March 23, 2020. Date of publication April 10, 2020; date of current version August 9, 2021. The work is supported in part by National Natural Science Foundation of China under Grants 81671766, 61971369, U19B2031,61671309, in part by Open Fund of Science and Technology on Automatic Target Recognition Laboratory 6142503190202, in part by Fundamental Research Funds for the Central Universities 20720180059, 20720190116, 20720200003, in part by Tencent Open Fund. The Associate Editor for this article was Y. Gao. *(Corresponding author: Yue Huang.)*

The authors are with the Key Laboratory of Underwater Acoustic Communication and Marine Information Technology, Ministry of Education, School of Informatics, Xiamen University, Fujian 361005, China (e-mail: yhuang2010@xmu.edu.cn).

Digital Object Identifier 10.1109/TITS.2020.2983475

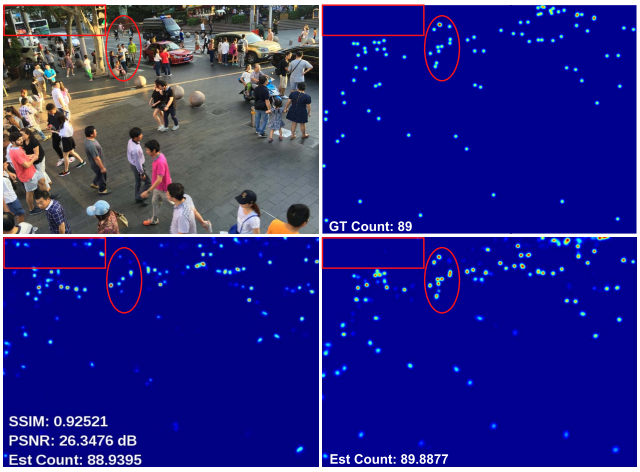

Fig. 1. Density estimation comparison. Top left: the original image (from Part_B of ShanghaiTech dataset [49]). Top right: the ground truth image. Bottom left: the images generated by CP-CNN [38] (direct screenshots from the CP-CNN). Bottom right: the image generated by our network. Red box indicates False Positive and red circle indicates False Negative.

with other methods such as global number estimation or crowd density level estimation and achieved significant improvements. The perspective effect and density change make a crowd counting task very challenged. To address this issue, FCN [25] proposed a multi-scale average prediction method. Recently presented ensemble methods such as MCNN [49], Hydra CNN [28], Crowdnet [3] and Switch-CNN [32] have the adaptability to scale changes by using a multi-column or divide-and-conquer strategy, but these methods solve the scaling problem only to some extent. Besides, the context-aware approaches such as that presented in [33] and a CP-CNN introduced in [38] incorporate global and local contextual information into a convolutional network to reduce an estimation error.

However, we noticed that CNN models with a complex structure do not deal with the multi-scale problem well enough, and an improvement is still required. Moreover, most existing methods focus only on the accuracy of crowd counting neglecting the correctness of density distribution. As illustrated in Fig. 1, if the focus is only on a total count, the CP-CNN [38] can estimate the count accurately, but there are obvious errors in the density distribution. Specifically, the resulting accuracies are close to the optimal ones because the number of false negative and the number of false positive are almost the same, which can be intuitively reflected by the red box and red circle in Fig. 1. Obviously, there is nothing in the red box, but the CP-CNN "thinks" there are many people. In contrast, there are many people in the red circle, but CP-CNN missed a lot of them. Besides, the Mean Absolute Error(MAE) is not enough to estimate the density distribution. Therefore we introduce the Patch Absolute Error(PAE) to enhance the estimation accuracy between the counts and density maps. The PAE is defined as a total absolute deviation of an image patch. The smaller the image patch is, the more accurately the PAE can evaluate the correctness of the count and rationality of a density map.

In crowd images (as show in Fig. 1), it is common for perspective distortion and occlusion problems. These disturbances make the size of the heads in the images change obviously, and some details of the head cannot be captured well. Hence, it is difficult for the traditional hand-crafted features algorithm to extract the detailed information and achieve effective performance in the dense crowd images. The CNN-based algorithms benefit from their powerful feature extraction capabilities, and can often learn more effective feature representation that the traditional hand-crafted method. It brings a large performance boost for dense crowd counting. Therefore, we also consider using CNN to achieve high accuracy in crowd counting.

Moreover, the existing approaches employ the pooling layers which results in a low-resolution and features loss. In [15], the author observed that deeper layers encode the high-level knowledge which includes a rich semantic information, while shallower layers capture the low-level features which include a rich spatial information. Obviously, combining the information from the shallow and deep layers is the best option. Therefore, we propose a novel symmetric CNN framework to reduce a count error and generate more a reasonable density map by making a full use of the multi-layer features. In our network we make full use of the information in both encoder and decoder stage. At the same time, we use several up-sampling operations to achieve a high-resolution density map of the input images. Such a network structure realizes an effective combination of multi-layer features, which can strengthen the feature propagation and encourage the feature reuse. The main contributions of this study can be summarized as follows:

- We propose a novel symmetric CNN architecture to train an end-to-end network that can predict a crowd density map by combining the features from different layers. This architecture can improve the feature utilization and reduce the number of false negatives and false positives in a generated density map. The extensive experiments are conducted on several major crowd counting datasets, which are challenging and representative. The experimental results demonstrate that our method achieves better performance compared with the state-of-the-art methods.
- We introduce a new evaluation PAE method to measure the quality of a generated density map. The PAE focuses not only on the accuracy of the crowd count, but also the rationality of the density map.

The paper is organized as follows. In Section II, we describe the proposed method in detail. In Section III, the obtained experimental results are presented and the corresponding analysis is provided. In Section IV, we give a brief conclusion.

II. Methodology

Compared with the regression methods whose output denotes the number of people, the regression methods whose output is a density map of an input image can provide more information for the crowd analysis. The distribution of density maps can be used to analyze the abnormal situation in the image. So, for a given image, the network we proposed here will output the density map of the image, and then get the number of the people by a integration procedure [49]. Therefore, we first describe how to generate

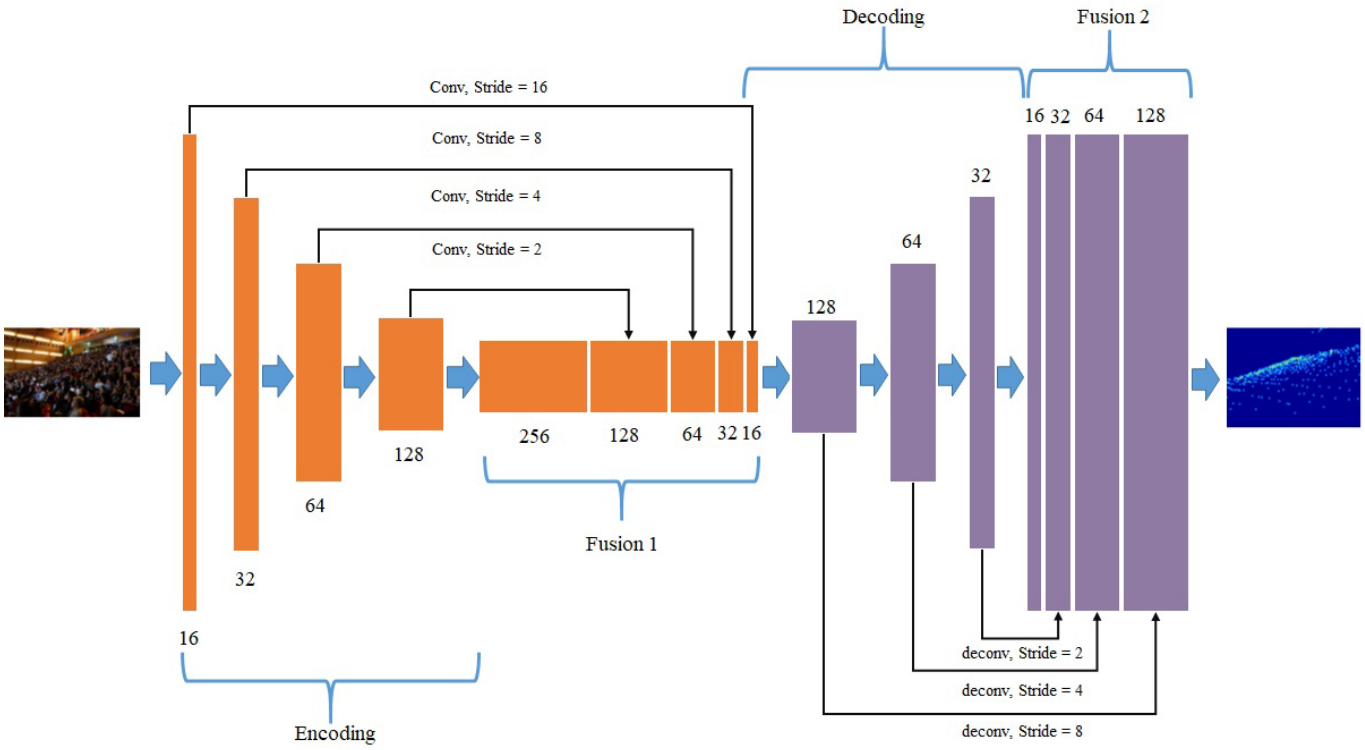

Fig. 2. The proposed symmetric architecture for crowd counting. Blocks in orange and purple denote encoding network, decoding network and two fusion networks respectively. (Better visualization in color version).

a crowd density map by using a head annotation of an image.

Compared with semantic segmentation, the labeling procedure in crowd counting is in a different manner. The labels are scatters that locate inside the targets. Due to complex human postures, correct labeling scatters is allowed to locate at any part of the target. So it is easily affected by the label makers. Different label makers may generate different labels for the same targets in surveillance images. In addition, there are some labeling mistakes in the high-density areas. Therefore, the proposed method addresses these unique challenges in crowd counting as follows:

A delta function $\delta(x - x_i)$ is used to denote a head at pixel x_i. Hence, we can represent an image with N heads labels as follows [49]:

$$H(x) = \sum_{i=1}^{n} \delta(x - x_i) \tag{1}$$

And then, a Gaussian kernel G_σ is used to describe the heads distribution in the image. With this Gaussian diffusion G, we assume that the center point has the greatest confidence and then gradually decreases in the surrounding area, thus covering the entire target as much as possible. So the density of a image is defined by convolving the image with the Gaussian kernel, as

$$F(x) = H(x) * G_\sigma \tag{2}$$

where $F(x)$ is the estimated density, $H(x)$ is an image with N heads derived from eq.(1). To simplify the implementation, we propose to use the fixed kernels G_σ instead of the geometry-adaptive kernels $G_\sigma(x)$ in [49].

A. Network Framework

In crowd counting, our main goal is to estimate the number of people in an image accurately. Given an image, the network will output the corresponding density map, and the number of people in the image can be obtained by integrating the density map. However, on one hand, due to the existence of a perspective distribution, the size of the heads in an image changes very rapidly. On the other hand, severe occlusions that appear makes it difficult to distinguish individuals. Therefore, the crowd counting is a very challenging task. Current state-of-the-art methods use multi-column architectures to solve the multi-scale problems [32], [38], [49]; they tend to map heads with the multi-scales by using the convolution kernels with different sizes. This multi-column framework is able to overcome the impact of different sizes of human heads caused by a perspective distribution up to a certain level. However, this structure needs to set a reasonable size of a convolution kernel for different columns. The rationality of a convolution kernel largely influences the network performance, making it too complicated because a multi-column architecture requires a lot of calculations. Recently, a great success in salient object detection [15] has been achieved by combining the information from different semantic layers. In [15], the author showed that shallower layers could provide rich spatial information, whereas the deeper layers encode a high-level semantic knowledge and can better determine the object distribution.

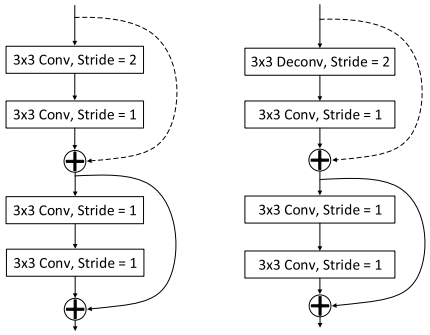

Fig. 3. Details of encoding block (left) and decoding block (right) in Fig.2. The two blocks are only little different in the first convolution.

Thus, by combining the features from both shallower and deeper layers can represent the object better.

Inspired by the existing works, we suggest that:

i) a good crowd counting network should combine the information from different layers;

ii) the shallow detail features combined with the deep semantic information can represent the features information on an object better.

Therefore we want to develop a new method for crowd counting by combining the features from different layers. Hence, we propose a symmetric encoder-decoder CNN architecture for crowd counting shown in Fig. 2. The proposed architecture is more detailed described as follows:

- **Encoding network**: It consists of five encoding blocks that represent different scale information, each encoding block is more detailed provided in Fig. 3 (left). Dotted lines indicate that 1×1 convolution with a stride of two is used to match the dimensions.
- **Fusion network 1**: As presented in the motivation that the proposed work aims to combine encoding information of different scales (Fig. 2). The convolution operation is use to down-sample the features to achieve dimensional matching. Here, fusion of the multi-scale and multi-semantic features is beneficial for solving the interference problem in crowd counting
- **Decoding network**: It consists of five decoding blocks that represent different features information, the decoding block is more detailed shown in the Fig. 3 (right). Dotted lines indicate that 1×1 convolution with a stride of two is used to match the dimensions.
- **Fusion network 2**: For different features information in the decoding procedure (Fig. 2), they represent characteristic information at different scales. So combining these features can generate more accurate estimation.

In the encoding stage, at each down-sampling step, the number of output channels is set such that to be twice greater than the number of the input channels to reserve more information. After four down-sampling operations, these feature maps represent only 1/16 of the original image. However, these feature maps have a larger receptive field than the previous feature maps. According to the the results presented in [15], the scale information of different coding layers is different, so the information of different layers should be combined to get a latter decoding network.

In the decoding stage, the network decodes the information from different receptive fields in turns. Different decoding layers can represent heads with different sizes, which have various semantic information. So the feature maps from different decoding layers should be merged. Finally, a 1×1 convolution operation is used to generate the last density map.

In order to extract effectively features, we combine the features from different scales and different semantics together. The feature fusion of different scales facilitates the extraction of multi-scale features, and the fusion of different semantics is beneficial for the network to encode the human heads information in different sizes and congestion levels in the image. Therefore, the feature fusion of different scales and different semantics can improve the final accuracy. The presented symmetric encoder-decoder network structure can bring the following three advantages:

- The network does not require a multi-column structure to learn the information from different receptive fields anymore. Only a single-column network can obtain the information from different receptive fields. So the complexity of network is reduced.
- Both encoding and decoding networks combine the information from different layers to reuse the information, and then to predict a more accurate estimation.
- Such a hourglass structure can reduce not only the memory consumption, but also the computational cost.

The architecture is similar as the ones in semantic segmentation. Here we want to claim that the proposed work is motivated by the unique challenges from crowd counting: For semantic segmentation, the authors usually consider that the outputs from shallower layers can capture rich spatial information. These low-level features benefit describing the details of the objects. Then the high-level features from deeper layers can be transformed to shallower layers for locating the region with higher accuracy. By combining features from different levels together, the framework can provide richer multi-scale features from each layer, and generate a satisfactory segmentation map. Hence, the purpose of feature aggregation for the pixel-level semantic segmentation task is to describe the more accurate spatial location and local details [2], [26], [31], [34]. However, in the proposed crowd counting task, we use filters with different size of receptive fields to extract features at different scales. That is because the size and the shape of target (e.g. head) in the surveillance images are inconsistent due to the perspective effect or image resolution. Hence, we need to combine the features from different scales together to generate a density maps. On the other hand, the U-Net [36] only used the features from the last layer, and the shortcut is used to provide more details. But in our crowd counting task, the framework should combine the features from different scales together to accommodate the targets of different sizes. So we combine the features in all encoding and decoding layers. In addition, in the decoding layers, each layer output generate a characteristic representation of the target head at a specific scale, and combining them would generate a better density map.

TABLE I

INFORMATION ON SIX DATASETS USED IN THE EXPERIMENTS. TOTAL IMAGES FIELD DENOTE THE SUM OF TRAIN IMAGES AND TEST IMAGES; RESOLUTION MEANS THE RESOLUTION OF THE IMAGES IN THIS DATASET; MIN, MAX AND AVE MEANS MINIMUM, MAXIMUM AND AVERAGE NUMBER OF PEOPLE IN A IMAGES, RESPECTIVELY

Dataset		Total images	Resolution	Min	Max	Ave	Total count	Who use
ShanghaiTech [49]	Part_A	482	Different	33	3139	501	241,677	[25], [32], [38], [46], [49]
	Part_B	716	768*1024	9	578	123	88,488	[32], [38], [46], [49]
WorldExpo'10 [46]		3980	576*720	1	253	50	199,923	[32], [38], [46], [49]
Mall [8]		2000	640*480	13	53	31	62,325	[20], [36], [42]
UCF_CC_50 [16]		50	Different	94	4543	1279	63,974	[10], [25], [49]
SmartCity [48]		50	1920*1080	1	14	7	369	[32], [38], [46], [49]
Beijing BRT [10]		1280	640*360	1	64	13	16795	[32], [48]

B. Loss Function

The existing methods [3], [8], [32], [38], [49] generally use the Euclidean distance to measure the error between a generated density map and the corresponding ground-truth. Here, we follow these line of research to define the loss function:

$$L(\Theta) = \frac{1}{2N}\sum_{i=1}^{N} ||F(X_i;\Theta) - F_i||_2^2 \tag{3}$$

where, $L(\Theta)$ denotes the overall losses, N is the number of training images, X_i represent the ith input image, Θ represents network parameters, $F(X_i;\Theta)$ represents the generated density map from X_i, and F_i is the ground-truth density map of X_i.

III. EXPERIMENTS AND EVALUATION

In this section, the evaluation of our network performance on several public crowd counting datasets: ShanghaiTech [49], WorldExpo'10 [46], Mall [8], UCF_CC_50 [16], SmartCity [48] and Beijing BRT [10], is presented. The detailed information on these datasets is shown in Table I. First, in order to demonstrate the effectiveness of different layers concatenation, an ablation study using the Part_A of ShanghaiTech dataset is presented. Then, the experimental results are provided to demonstrate the rationality of using a fixed kernel to generate the density map. Afterwards, the proposed method is compared with the current state-of-the-art methods on above-mentioned datasets (different dataset has different characteristic, so we select these datasets evaluation our method. We should know due to there is no public code for most above-mentioned methods, so we use the result there had shown in paper). Lastly, the result of a few experiments are illustrated the to prove effectiveness of a new evaluation method. In the comparison, we used the Mean Absolute Error (MAE) and Mean Squared Error (MSE) to measure the count error, which were defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|z_i - \hat{z}_i|, MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i - \hat{z}_i)^2} \tag{4}$$

where, N is the total number of images in test dataset, z_i is the annotated number of the ith image, and $\hat{z}_i$ is the estimated number of people. Generally speaking, MAE shows the accuracy of our model in crowd counting, while MSE means the robustness of our model in crowd counting.

In training, four down-sampling and up-sampling operations were performed in our network. For convenience, the size of the input images patch was a multiple of 16. In the experiments, the train patches with the size of 160 × 160 were randomly cropped from the original images. Horizontal flipping with the probability of 0.5 was used for data augmentation. When the network performance was tested, the original images were directly sent to the model (when the length or width of the images was not appropriate, zero will be used in the right or bottom part of the images, making the length and width of the images a multiple of 16). The stochastic gradient descent (SGD) was selected to minimize $L(\Theta)$, the momentum was set to 0.9, and the weight decay was set to 10^{-5}. Our network was based on Caffe framework presented in [18]. The server parameters were as follows: Intel(R) Xeon(R) CPU E5-2683 v3 @2.00GHz, 128G RAM and NVIDIA GeForce GTX1080 Ti.

A. Ablation Study

In this section, an ablation study is presented to verify the effectiveness of combining the information from different layers. Our network performance was verified using four groups of experimental setups: (1) The experiment without concatenation: there was no concatenation of information from different layers between encoding and decoding stage, we use it as our base network.(2) The experimental with encoding concatenation: the information from different layers was concatenated only in the encoding stages. (3) The experimental with decoding concatenation: the information from different layers was concatenated only in the decoding stage. (4) The experimental with concatenation both encoding and decoding stages(the proposed network): the information from different layers was concatenated both in encoding and decoding stages respectively.

All of our comparison experiments were performed using the Part_A of ShanghaiTech dataset [49]. Since this dataset contained a lot of scenes, the head size of people in the images changed dramatically, and images in this dataset included both relatively sparse and seriously dense crowd. The Part_A, first introduced in [49], consisted 482 images collected from the Internet.

The detailed comparison of experimental results is shown in Table II. Wherein it can be see that the experiments result with the concatenation in the encoding stage or decoding stage had a slight improvement over the experiments result without concatenation. The best results were achieved when

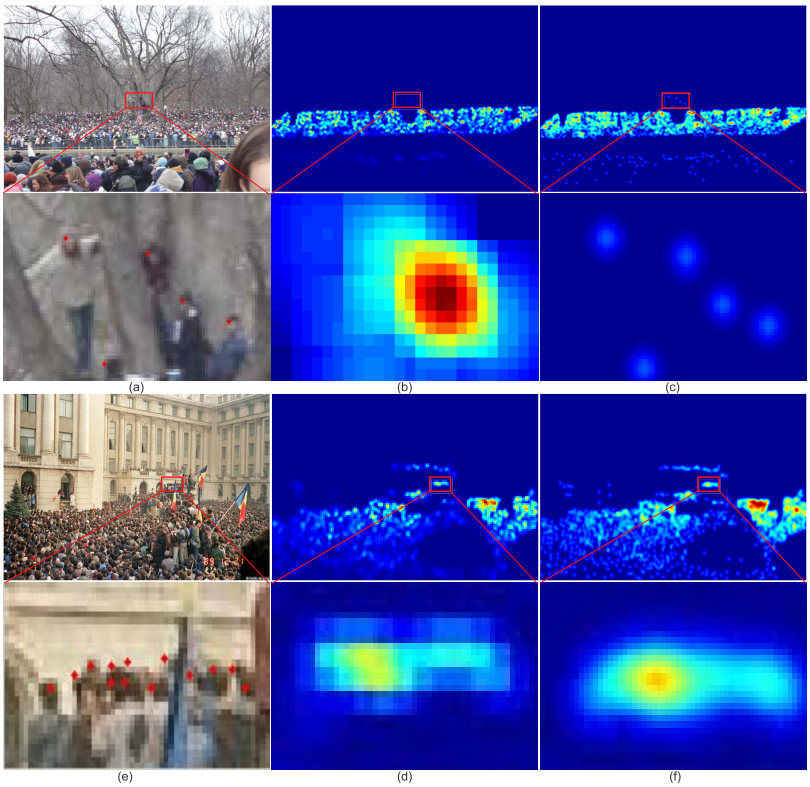

Fig. 4. Rationality of a fixed kernel. (a-b) the original image in ShanghaiTech Part_A [49], (c-d) the density maps corresponding to (a-b) respectively generated by the geometry-adaptive kernels, respectively, (e-f) the density maps corresponds to (a-b) generated by the fixed kernels, respectively; red box denotes the density map is not suitably generated by the geometry-adaptive kernels. In (b), to make the image look more intuitive, the contrast is changed.

TABLE II
RESULTS OF THE ABLATION STUDY. BOLD DENOTE THE BEST RESULTS

	MAE	MSE
Without concatenation	77.9	128.3
Encoding concatenation	76.5	122.7
Decoding concatenation	76.2	120.0
The proposed method	**69.8**	**114.7**

the concatenation was used in both stages. The main reason for such result may be that a fusion of multi-layer feature in both encoding and decoding stages can better represent the head of people, since the information between encoding and decoding stages is complementary, a fusion of multi-layers feature in only encoding or decoding stage is not good enough to represent the head. Therefore, a slight improvement in concatenating in encoding or decoding stage was achieved. The further improvement was achieved when a concatenation of information from different layers was implemented in both encoding and decoding stages. These results prove the effectiveness of combining the information from different layers.

B. Rationality of Fixed Kernel

In [21], a density map was used in crowd counting for the first time. After that, the benefits of using a density map in crowd counting has been recognized. Namely, the density map can preserve more spatial distribution information, which is beneficial for crowd behavior analysis. In an MCNN [49], the author showed the head size is related to the distance between the centers of k neighboring persons in crowded counting. Thus, the average distance between nearest k neighbors person can reasonably estimate the current head size (they call this method the geometry-adaptive kernels). In this way, the dependency on the perspective map is solved, but this method can be applied only to the dense scene with a relatively large head. For a sparse scene or smaller head size in the image, the kernel size will become either too big or too small.

The rationality of a fixed kernel is presented in Fig. 4. In Fig. 4 (b), it can be see that the geometry-adaptive kernels use a large area representing a small person in the distance, while in Fig. 4 (d), the geometry-adaptive kernels use only several pixels representing a small person in the distance. Specifically, when the pooling operation is repeatedly used in the network (such as MCNN, CrowdNet, DR-ResNet, Switch-CNN), the output density map is small. In order to make the generated density map correspond to the output density map, we used only one pixel to represent a human head. Thus, regression to the density map was degraded to the regression to counting, which was difficult for our network. Compare to Figs. 4(b) and (d), the density map in Figs. 4(c) and (f) seem more reasonable, wherein the head size is represented by a fixed size.

TABLE III

NETWORK PERFORMANCE USING THE FIXED KERNELS AND GEOMETRY-ADAPTIVE KERNELS

	Part_A		UCF_CC_50	
	MAE	MSE	MAE	MSE
geometry-adaptive kernels	68.1	109.3	274.0	380.4
fixed kernels	69.8	114.7	271.3	376.3

TABLE IV

PERFORMANCE COMPARISON ON SHANGHAITECH DATASET

	Part_A		Part_B	
Method	MAE	MSE	MAE	MSE
Zhang et al. [46]	181.8	277.7	32.0	49.8
MCNN [49]	110.2	173.2	26.4	41.3
FCN [25]	126.5	173.5	23.8	33.1
Switch-CNN [32]	90.4	135.0	21.6	33.4
CP-CNN [38]	73.6	**106.4**	20.1	30.1
DR-ResNet [10]	86.3	124.2	14.5	21.0
Ours	**69.8**	114.7	**10.2**	**14.9**

At the same time, due to the limited perspective map, we could not accurately estimate the perspective distribution of the images. Specifically, in practice, the perspective distribution depends on the camera installation height, angel, position, focal length, etc. Determining the perspective distribution of each camera is an enormous task.

Taking into account the above problem, as a compromise, we propose to use the fixed kernels instead of the geometry-adaptive kernels. The density maps in Figs. 4 (c) and (f) seem more reasonable than those presented Figs. 4 (b) and (d). An experiment was conducted to demonstrate the rationality of the fixed kernels. Experimental results for both geometry-adaptive kernels and fixed kernels are presented in Table III, where it can be noticed that the achieved performances were comparable, but for the fixed kernels, the density map was easily generated, and it was indented of k and perception map. The main reason for that is that density map was a fuzzy representation of the head, so the simple fixed kernels were enough.

C. Evaluation and Comparison

1) ShanghaiTech Dataset: The Part_A of the ShanghaiTech dataset was introduced in Section III.A. The images in Part_B were collected from the busy, urban streets in Shanghai. Due to a lot of dense crowds in the ShanghaiTech dataset, it was a very challenging task to estimate the number of people in the image accurately. So we used the ShanghaiTech dataset to evaluate the performance of our network in the dense crowds. The Ground truth density maps were generated using the fixed kernels (for more details, please refer to Section III.B). We compared our method with the recent state-of-the-art methods.

Table IV shows that our network achieved the lowest MAE and a relatively comparable MSE on the extremely dense crowds with varying density levels and perspective distortion. In Part_B, the results of our method were significantly improved compared to other methods. In addition, the example of Part_B is shown in Fig. 1, where it can be seen that our network could accurately locate the distribution of people. This indicates that our network can clearly distinguish the difference between people and background in a complex scene.

TABLE V

PERFORMANCE COMPARISON ON WORLDEXPO DATASET

Method	Scene1	Scene2	Scene3	Scene4	Scene5	Average
Zhang et al. [46]	9.8	14.1	14.3	22.2	3.7	12.9
MCNN [49]	3.4	20.6	12.9	13.0	8.1	11.6
Switch-CNN [32]	4.4	15.7	10.0	11.0	5.9	9.4
CP-CNN [38]	2.9	14.7	10.5	**10.4**	5.8	8.86
Ours	**2.8**	**12.1**	**9.4**	15.6	**3.5**	**8.68**

2) WorldExpo'10 Dataset: The WorldExpoİÂ¯10 dataset is a cross-scene crowd counting dataset, which was introduced in [46]. The dataset consists of 3980 video frames captured from 108 surveillance cameras in Shanghai 2010 WorldExpo. In [46], the authors provided different regions of interest (ROI) for every surveillance cameras and perspective maps for all scenes. Due to that, this dataset was used to measure the capacity of our network in a cross-scene crowd counting.

However, the some datasets settings as previously and the fixed kernels were used(for more details, please refer to Section III.B). For this dataset, we considered only the areas inside the ROI, so the output of the last convolution layer was modified based on the ROI mask. The MAE was calculated separately for each test scene to show how well the network fitted the current scene. The average MAE indicated the overall performance of the network in five test scenes. Our method was compared with four recent state-of-the-art methods, and obtained results are present in Table V. The results show that our network achieved the best average MAE. In the individual scenes comparison, our result are not really good in the Scene4, we think it may be because the images of this scene are very blurry. This experimental result shows that our multi-scale and multi-semantic network still need to be improve for the blurred images. But overall, our method is still competitive and can achieve the best performance in the most of the scenarios. These experimental results prove that our network has good cross-scene generalization performance.

3) Mall Dataset: The mall dataset contained 2000 images chosen from one surveillance camera in a shopping mall. Following the principle used in [8], we used frames 1-800 as the training set, the other frames as the testing dataset. We compared the results of our method with three most recent methods: CNN-boosting [42], MoCNN [20] and Weight V-LAD [36]. Table VI shows that our method achieved the best MAE and MSE by a significant margin. This indicates that our model can also estimate images from a single scene with relative sparse people.

4) Beijing BRT Dataset: The Beijing BRT dataset contained 1280 images collected from video surveillance at Beijing Bus Rapid Transit station; all the images were high-angle shots. Following the setting in [10], 720 images were select for training, and the remaining images were selected for testing. These images contained the shadows, glare, and sunshine interference, and time span was from morning till night; these factors commonly appear in real life, so this dataset can

TABLE VI
PERFORMANCE COMPARISON ON MALL DATASET

Method	MAE	MSE
CNN-Boosting [42]	2.01	-
MoCNN [20]	2.75	12.4
Weight V-LAD [36]	2.41	9.12
Ours	**1.85**	**2.34**

TABLE VII
PERFORMANCE COMPARISON ON BEIJING BRT DATASET

	MAE	MSE
MCNN [49]	2.2	3.4
FCN [25]	1.7	2.4
DR-ResNet [10]	1.4	2.0
Ours	**1.4**	**2.0**

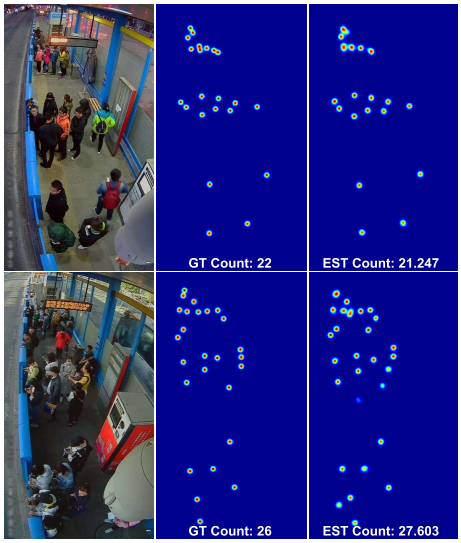

Fig. 5. Results of the proposed method on Beijing BRT dataset.

better reflect the performance of our network in practice. Further, we used the Beijing BRT dataset for evaluation of generalization performance of the model in daily life.

Table VII shows the comparison of results of MCNN, FCN, DR-ResNet, and our method. Our method achieved the best MAE of 1.36, and comparable MSE of 2.02 on Beijing BRT dataset, which proved that our model could be applied in daily life. The example is shown in Fig. 5.

5) UCF_CC_50 Dataset: The UCF_CC_50 dataset [16] contained 50 images collected from the internet. The limited number of the images and a large variance in the crowd count, make it a very challenging dataset. Moreover, the head size in this dataset was usually less than 10 pixels, which also increased the difficulty of crowd counting. For a fair comparison, we perform 5-fold cross-validation according to the stand setting [16]. We compared the results of our method with those of recent state-of-the-art methods, and the comparison is presented in Table VIII, where it can be seen that our

TABLE VIII
PERFORMANCE COMPARISON ON UCF_CC_50 DATASET. TRAINING IS PERFORMED USING THE PART_A, SO OUR MODEL WAS TRAINED IN THE SOURCE DOMAIN

Method	MAE	MSE
Zhang et al. [46]	467	498.5
MCNN [49]	377.6	509.1
Switching-CNN [32]	318.1	439.2
CP-CNN [38]	295.8	**320.9**
DR-ResNet [10]	307.4	421.6
Ours	309.1	428.8
MCNN (Transfer learning) [49]	295.1	490.2
Ours(Finetune the whole network)	**271.3**	376.3

network achieved a competitive MAE compared with the best performance realized by the CP-CNN (309.1 vs 295.8).

Futher, to demonstrate the generalization performance of our network, by following the principle in [49], we evaluated our network performance in transfer learning with extremely dense crowd images. Since Part_A was relatively complex compared to other datasets, we set Part_A of ShanghaiTech dataset as a source domain, while UCF_CC_50 dataset was set as a target domain. The UCF_CC_50 dataset was only used to fine-tune the network trained using the Part_A dataset. The results are shown in the last line in Table VIII. After network fine-tuning, our network achieved the lowest MAE and comparable MSE. This result demonstrates that our network has good generalization performance in transfer learning.

6) SmartCity Dataset: The SmartCity dataset was proposed in [48], and it consists of relatively sparse pedestrians both indoor and outdoor (more detail information on SmartCity is presented in Table I.); all the images in this dataset were collected from video surveillance of ten city scenes including the office entrance, sidewalk, atrium, shopping mall, etc. This dataset is different from the existing crowd counting datasets because only sparse pedestrians appear in every image. This dataset was also used to evaluate our network performance in transfer learning, which can demonstrate the generalization of our model. For a fair comparison, we followed the transfer learning settings used in [48]. We used the Part_B of ShanghaiTech dataset as a source domain, and the SmartCty dataset as a target domain; our network was trained in a source domain; and then tested it in a target domain without fine-tuning.

In Table IX, the comparison of the results of MCNN [46], Switch-CNN [32], SaCNN [48], and our model is presented. All models were only trained on Part_B dataset. Table IX shows our model had the lowest MAE of 5.9, these results have a significantly marginal compared with the SaCNN. The presented results clearly reflect the generalization ability of our network. Consequently, according to the results on all used datasets, it can be concluded that our model had the best performance on very sparse crowd scenes among all tested methods.

D. New Evaluation Standard for Crowd Counting

The MAE is used to evaluate the accuracy of estimation of crowd counting methods, and MSE is used to evaluate

TABLE IX

PERFORMANCE COMPARISON ON SMARTCITY DATASET. SACNN(W/O CL) DENOTE THE SACNN WITHOUT A COUNT LOSS

	MAE	MSE
MCNN [49]	40.0	46.2
Switch-CNN [32]	23.4	25.2
SaCNN(w/o cl) [48]	17.8	23.4
SaCNN [48]	8.6	11.6
Ours	**5.9**	**10.4**

the robustness of the estimation. However, the MAE only shows whether the total numbers estimated by a generated map and the ground truth map are the same, ignoring the rationality of a generated map. In CP-CNN, the author used SSIM to evaluation the quality of a density map, this is first one that focus on the quality of density maps. But SSIM is compared pixel by pixel, it is not suitable for crowd counting. In order to evaluate the accuracy and reasonability of a density map, we introduce a new evaluation standard named the PAE. Unlike the MAE, the PAE calculates the deviations of the number of people in each patch from a density map. The PAE is obtained by summing the absolute value of all deviations in a density map. So the PAE will consider the location information in a image. We can also use PAE to evaluate the performance of the model on a dataset, here we call it PAE-A. Formally, PAE-A is the average of the image PAE. Using this novel method, we can determine the number of the false negatives and false negatives in a density map which are ignored in the MAE. Thus, the combination of PAE and MAE can give a better insight into a density map's rationality. The PAE is defined by:

$$PAE = \sum_{i=1}^{N_{patch}} |p_i - \hat{p}_i| \tag{5}$$

where N_{patch} denotes the number of patches in the image (the area of each patch is equal, and here N_{patch} is equal to 16), p_i is the ground truth of people in the ith patch, and $\hat{p}_i$ is the estimated number of people in the ith patch. For PAE, the closer to MAE, the more accurate the model.

To explain a new evaluation method more clearly, we present an example in Fig. 6. The MCNN achieved a precise count result but the focus was only on the MAE. However, when the density map generated by MCNN was analyzed, the false negatives and false positives both appeared in a density map. In contrast to the density map presented in Fig. 6 (d) which was generated by our network, although the estimated number of people was relatively mall, we produced a more reasonable and more correct density map. In order to explain this phenomenon more objectively, we divided the density map into N_{patch}=16 patches (we could also divide it into smaller pathes, this is an experience value, generally choose according to the density of the crowd. In our experiment we found N_{patch}=16 patches were enough to evaluate the rationality of a density map), and there was no overlap between them. Then, we calculated the deviations in the patch. The PAE was calculated by summing the absolute deviations.

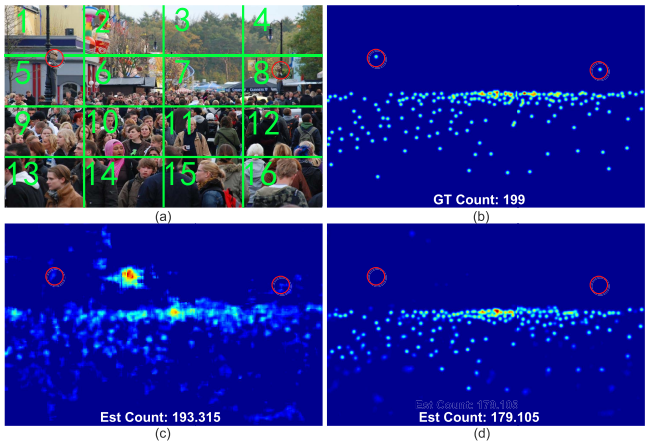

Fig. 6. Density estimation comparison. (a): The Original image (from the Part_A of the ShanghaiTech dataset [49]). (b): The Ground truth. (c): The image generated by the MCNN [49]. (d): Generated by our network. Red circle indicates puppets appeared in this image.

The detailed comparison results are shown in Table X. Tough, there are larger deviations in the total number (19.90 vs 5.69), the lower PAE than that of MCNN was achieved (25.97 vs 59.08. The PAE was calculated only for a specific image. when it was calculated on the whole dataset, it became 91.18 vs 144.26); also there were not larger deviations in each patch. In contrast, the MCNN had lager deviations in patches 2, 6, 7 and 9. This demonstrates the rationality of the density map generated by our network, and it is consistent with the result of PAE.

We also give a comparison of our method and MCNN based on the ShanghaiTech Part_A and Part_B datasets, and use PAE-A(PAE-A is the average of the image PAE) as the evaluation criteria. The experimental results are shown in Table XI.

From Table XI, it can be found that there is a large false positive and false negative in the model. Especially, Our model has smaller deviations both in Part_A(32.9 vs 21.4) and Part_B(5.8 vs 4.2). This shows that PAE can measure false positive and false negative in the image, and our results are also more reliable. One point to note about the results of the above table is that when N_{patch} is 1, then PAE-A is the same as MAE. When N_{patch} is the same as the total pixels of the image, PAE becomes a pixel-by-pixel metric, and it is the similarity as PSNR and SSIM. But PSNR and SSIM sensitive to pixel deviations, it is not suitable for crowd counting. Due to the label data has pixel-level deviations, and this deviation does not affect the result of the final count. From this perspective, PAE is actually a trade-off evaluation method between MAE and SSIM, it can measure the rationality of the density map more clearly.

In the following paragraph, some interesting phenomenon in our experiment is presented, which even better illustrates the performance of our model.

The red circle in Fig. 6 (a) includes the puppets. However, there is still annotated information above. In Figs. 6 (c) and (d), the density maps generated by the MCNN and our model are presented. It's seem to our method have a poor performance,

TABLE X

THE COMPARED OF THE PROPOSED METHOD COMPARISON AND MAE IN PART_A. DUE TO THE LOSS OF CALCULATION ACCURACY, THERE ARE SOME DEVIATIONS IN SUM COMPARE WITH THE LABEL IN FIG. 6

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	SUM	MAE
Ground Truth	0.16	0	0	0	15.03	33.64	54.83	25.88	23.61	21.23	9.43	8.99	2.00	1.01	2.00	1.00	198.81	-
MCNN [49]	1.52	10.82	0.17	0.02	13.97	42.08	39.98	21.66	14.98	20.99	8.34	6.79	4.05	1.01	2.73	4.20	193.31	-
Ours	0	0.75	0	0.01	12.74	32.71	48.45	20.27	17.46	18.99	10.67	9.36	1.92	1.18	1.90	1.45	178.86	-
MCNN's deviations	+1.36	+10.82	+0.17	+0.02	-1.06	+8.44	-14.85	-4.22	-8.63	-0.24	-1.09	-2.20	+2.05	0	+0.73	+3.20	59.08	5.69
Our deviations	-0.16	-0.75	0	+0.01	-2.29	+0.07	-6.38	-5.61	-6.15	-2.24	+1.24	+0.37	-0.08	+0.17	-0.10	+0.45	25.97	19.90

TABLE XI

DEMONSTRATES THE RATIONALITY OF THE DENSITY MAP

	Part_A			Part_B		
Method	MAE	PAE-A	deviations	MAE	PAE-A	deviations
MCNN [49]	108.6	141.5	32.9	28.6	34.4	5.8
our	69.8	91.2	21.4	10.9	15.1	4.2

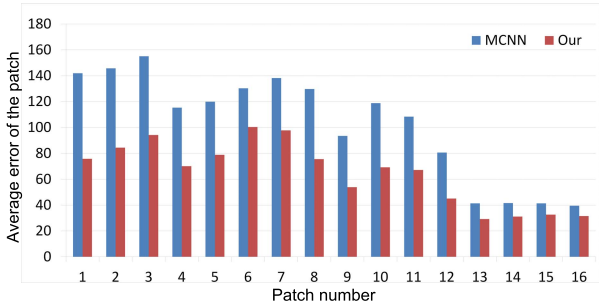

Fig. 7. The average error of the patch in Part_A. Patch number is defined as presented in Sec III.D.

but when we analysis the image careful. Which clearly shows that our model identified the dummy well, while MCNN failed. We must highlight that we did not purposed make the model learn from the real people or puppets, but our model could still identify them accurately. We assume that this is because our model combined information from different layers which helped describe the real people better.

Accordingly, the patch-based crowd counting not only could evaluate the rationality of a density map but also could determine the distribution of error position. However, our model had a poor estimation performance in some areas. So we calculated the average error of the patches in Part_A. First, all the images were divided into N_{patch}=16 patches, and then the average error was calculated at the corresponding patch position. The experimental results are presented in Fig. 7. Our network achieved the lower average errors in each patch, which demonstrated that our network have a good counting accuracy in each patch. Also, the error of two models was mainly concentrated in patches 1-12. Therefore, the perspective effects were common in images, resulting the head size became very small, and the crowd was very dense in the distance. This situation was very hard for the network to process, so the large deviations appeared in these areas.

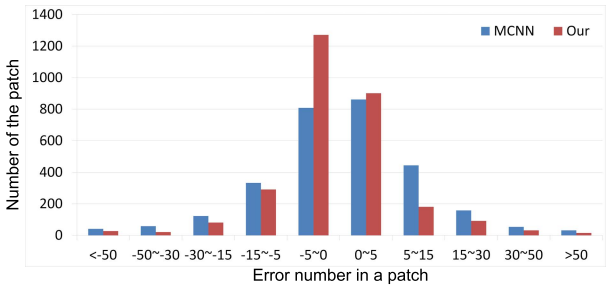

Fig. 8. Deviations number distribution of the Part_A.

Additionally, in order to analyze the deviations distribution of the patch, we collected and counted the deviations of all patches generated by our model and MCNN using the Part_A of the ShanghaiTech dataset. The results are shown in Fig. 8, where it can be seen that deviation was similar to a Gaussian distribution; especially, the deviations of the patches was mainly between -5 and 5 in our model, and only a few patches had larger deviations. We know that the range of -5 to 5 represents a small deviation, and that more image patches distributions within a small deviation range indicate that our network achieves an accurate crowd counting for most of the image patches. In contrast, it can be noticed that more deviations appeared in the ranges of less than -5 and greater than 5. This indicates the density map generated by our model had less false negatives and false positives. It can also be said that our results are more accurate. Hence, a more reasonable density map was generated, and a more reliable count was achieved by our method.

IV. CONCLUSION

We propose a novel symmetrical CNN to fuse the multi-layer features for crowd counting. The network automatically combines features from different layers in encoding and decoding stages to generate a more reasonable density map for achieving more accurate counting. Several major crowd counting datasets were used to evaluate the performance of our network in the single-, complex-, cross-scene, and to verity its ability of transfer learning. The extensive experimental results on mentioned datasets demonstrate that our network can achieve the state-of-the-art performance with a relatively simple network. Further, the PAE that assists the MAE to give a more reasonable evaluation of a density map quality is introduced for the first time. The experimental result show that the combination of PAE and MAE represents a more precise way to evaluate the rationality of density maps.

REFERENCES

[1] A. Albiol, I. Mora, and V. Naranjo, "Real-time high density people counter using morphological tools," *IEEE Trans. Intell. Transp. Syst.*, vol. 2, no. 4, pp. 204–218, Dec. 2001.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[3] L. Boominathan, S. S. S. Kruthiventi, and R. V. Babu, "CrowdNet: A deep convolutional network for dense crowd counting," in *Proc. ACM Multimedia Conf. (MM)*, 2016, pp. 640–644.

[4] K. Cao, Y. Chen, D. Stuart, and D. Yue, "Cyber-physical modeling and control of crowd of pedestrians: A review and new framework," *IEEE/CAA J. Automatica Sinica*, vol. 2, no. 3, pp. 334–344, Jul. 2015.

[5] A. B. Chan, Z.-S. J. 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.

[6] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2467–2474.

[7] K. Chen and J.-K. Kämäräinen, "Pedestrian density analysis in public scenes with spatiotemporal tensor features," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 1968–1977, Jul. 2016.

[8] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *Proc. BMVC*, vol. 1, 2012, p. 3.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[10] X. Ding, Z. Lin, F. He, Y. Wang, and Y. Huang, "A deeply-recursive convolutional network for crowd counting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1942–1946.

[11] M. Fu, P. Xu, X. Li, Q. Liu, M. Ye, and C. Zhu, "Fast crowd density estimation with convolutional neural networks," *Eng. Appl. Artif. Intell.*, vol. 43, pp. 81–88, Aug. 2015.

[12] G. Gan and J. Cheng, "Pedestrian detection based on HOG-LBP feature," in *Proc. 7th Int. Conf. Comput. Intell. Secur.*, Dec. 2011, pp. 1184–1187.

[13] S. Gao, M. Zhou, Y. Wang, J. Cheng, H. Yachi, and J. Wang, "Dendritic neuron model with effective learning algorithms for classification, approximation, and prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 601–614, Feb. 2019.

[14] G. L. Hamza-Lup, K. A. Hua, M. Le, and R. Peng, "Dynamic plan generation and real-time management techniques for traffic evacuation," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 4, pp. 615–624, Dec. 2008.

[15] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5300–5309.

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

[17] H. Idrees, K. Soomro, and M. Shah, "Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 1986–1998, Oct. 2015.

[18] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[19] A. Kowcika and S. Sridhar, "A literature study on crowd (people) counting with the help of surveillance videos," *Int. J. Innov. Technol. Res.*, vol. 3, no. 4, pp. 2353–2361, 2015.

[20] S. Kumagai, K. Hotta, and T. Kurita, "Mixture of counting CNNs: Adaptive integration of CNNs specialized to specific appearance for crowd counting," 2017, *arXiv:1703.09393*. [Online]. Available: http://arxiv.org/abs/1703.09393

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

[22] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[23] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in *Proc. Int. Conf. Image Process.*, vol. 1, 2002, p. 1.

[24] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis. (ICCV)*, Sep. 1999, pp. 1150–1157.

[25] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "Fully convolutional crowd counting on highly congested scenes," 2016, *arXiv:1612.00220*. [Online]. Available: http://arxiv.org/abs/1612.00220

[26] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[27] L. Oliveira and U. Nunes, "Pedestrian detection based on LIDAR-driven sliding window and relational parts-based detection," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2013, pp. 328–333.

[28] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 615–629.

[29] E. Principi, D. Rossetti, S. Squartini, and F. Piazza, "Unsupervised electric motor fault detection by using deep autoencoders," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 2, pp. 441–451, Mar. 2019.

[30] Z. Ren, K. Qian, Z. Zhang, V. Pandit, A. Baird, and B. Schuller, "Deep scalogram representations for acoustic scene classification," *IEEE/CAA J. Automatica Sinica*, vol. 5, no. 3, pp. 662–669, May 2018.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[32] 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)*, vol. 1, Jul. 2017, p. 6.

[33] C. Shang, H. Ai, and B. Bai, "End-to-end crowd counting via joint learning local and global count," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1215–1219.

[34] 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. 640–651.

[35] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, "Crowd counting via adversarial cross-scale consistency pursuit," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5245–5254.

[36] B. Sheng, C. Shen, G. Lin, J. Li, W. Yang, and C. Sun, "Crowd counting via weighted VLAD on a dense attribute feature map," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1788–1797, Aug. 2018.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[38] 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.

[39] 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.

[40] V. B. Subburaman, A. Descamps, and C. Carincotte, "Counting people in the crowd using a generic head detector," in *Proc. IEEE 9th Int. Conf. Adv. Video Signal-Based Surveill. (AVSS)*, Sep. 2012, pp. 470–475.

[41] M. Teichmann, M. Weber, M. Zollner, R. Cipolla, and R. Urtasun, "MultiNet: Real-time joint semantic reasoning for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1013–1020.

[42] E. Walach and L. Wolf, "Learning to count with CNN boosting," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 660–676.

[43] 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)*, 2015, pp. 1299–1302.

[44] X. Wu, G. Liang, K. K. Lee, and Y. Xu, "Crowd density estimation using texture analysis and learning," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, Dec. 2006, pp. 214–219.

[45] J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan, "IPrivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 5, pp. 1005–1016, May 2017.

[46] 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.

[47] J. Zhang, B. Tan, F. Sha, and L. He, "Predicting pedestrian counts in crowded scenes with rich and high-dimensional features," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1037–1046, Dec. 2011.

[48] L. Zhang, M. Shi, and Q. Chen, "Crowd counting via scale-adaptive convolutional neural network," 2017, *arXiv:1711.04433*. [Online]. Available: http://arxiv.org/abs/1711.04433

[49] 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.
[50] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 94–108.

Xinghao Ding (Member, IEEE) was born in Hefei, China, in 1977. He received the B.S. and Ph.D. degrees from the Hefei University of Technology, Hefei, in 1998 and 2003, respectively. From 2009 to 2011, he was a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, Pratt School of Engineering, Duke University, Durham, NC, USA. Since 2011, he has been a Professor with the Department of Communication Engineering, School of Information Science and Engineering, Xiamen University, Xiamen, China. His main research interests include image processing and machine learning.

Fujin He received the B.S. degree from Zhejiang Normal University, Zhejiang, China, in 2016. He is currently pursuing the master's degree with the School of Informatics, Xiamen University. His main research interests include machine learning and image processing.

Zhirui Lin received the B.S. and M.S. degrees from Xiamen University, Xiamen, China, in 2015 and 2018, respectively. His main research interests include machine learning and image processing.

Yu Wang received the B.S. degree from Fuzhou University, Fuzhou, China, in 2002, and the M.B.A. degree from Xiamen University, Xiamen, China, in 2014, where he is currently pursuing the Ph.D. degree with the School of Informatics. His main research interests include machine learning, embedded systems, and deep model compression.

Huimin Guo received the B.S. degree from Zhejiang Normal University, Zhejiang, China, in 2018. She is currently pursuing the master's degree with the School of Informatics, Xiamen University. Her main research interests include machine learning and image processing.

Yue Huang received the B.S. degree from Xiamen University, Xiamen, China, in 2005, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2010. She was a Visiting Scholar with Carnegie Mellon University from 2015 to 2016. She is currently an Associate Professor with the Department of Communication Engineering, School of Information Science and Engineering, Xiamen University. Her main research interests include machine learning and image processing.