CROWD COUNTING VIA MULTI-VIEW SCALE AGGREGATION NETWORKS

*Zhilin Qiu, Lingbo Liu, Guanbin Li, Qing Wang, Nong Xiao, Liang Lin**

School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

ABSTRACT

Crowd counting, aiming at estimating the total number of people in unconstrained crowded scenes, has increasingly received attention. But it is greatly challenged by the huge variation in people scale. In this paper, we propose a novel Multi-View Scale Aggregation Network (MVSAN), which handle the scale variation from feature, input and criterion view comprehensively. Firstly, we design a simple but effective Multi-Scale Feature Encoder, which exploits dilated convolution layers with various dilation rates to improve the representation ability and scale diversity of features. Secondly, we feed multiple scales of input images into networks to generate high-quality density maps in a coarse-to-fine manner. Finally, we propose a Multi-Scale Structural Similarity loss to force our networks to learn the local correlation of density maps. Extensive experiments on two standard benchmarks show that the proposed method can generate high-quality crowd density map and accurate count estimation, outperforming the state-of-the-art methods with a large margin.

***Index Terms*—** Crowd Counting, Multi-Scale Feature, Input View, Feature View, Criterion View

1. INTRODUCTION

With the rapid growth of the urban population, public safety has become a great challenge in city management. Most safety control measures relied on crowd counting, which estimates the crowd number from images and surveillance videos. However, the large scale variance of people from massive street images from social networks and real-time surveillance, is still one of the main obstacles for accurate estimation.

However, precisely estimating the count of crowd is extremely difficult. The major challenge is how to handle the huge variation of people scale. As shown in Figure 1, the scales of people range from a tiny dot to hundreds of pixels. Rather than directly regress the total number of people, the current methods acquire the count estimation by generating the crowd density map from the given image, which contains the spatial distribution information of crowd and is more meaningful for proceeding applications.

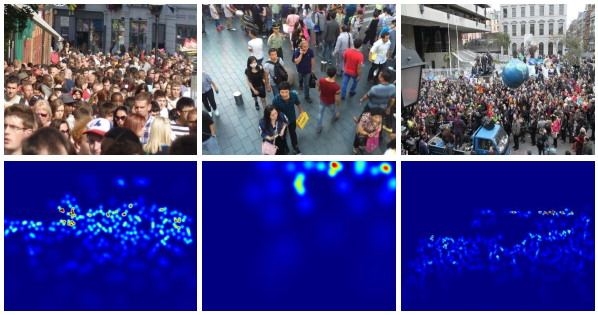

Fig. 1: Visualization of people with various scales in unconstrained crowded images from ShanghaiTech [1] dataset. Corresponding density maps are visualized in the bottom line. The huge scale variation of people is a major challenge limiting the performance of crowd counting.

With recent advanced deep convolutional neural networks(CNN), numerous network architectures have been proposed to address the task of crowd counting and have achieved remarkable performance [1, 2]. Among these methods, multi-scale architectures are the mainstream and most of them handle the scale variation of people with a multi-path network with different convolution kernel sizes on different paths. However, these methods still suffer from the following issues. Firstly, the limited paths restrict the capacity of multi-scale representation learning. Aimlessly increasing the number of paths is not worthy, since the optimization of mass integral parameters of these paths could fail on diversified scales feature learning, as revealed in [3]. Secondly, these networks contain multiple pooling layers and generate the low-resolution density maps, which are too coarse to estimate the accurate crowd count. Moreover, their insufficient exploration of pixels relation on the density maps leads to criticized blurring density maps.

To address the aforementioned drawbacks of current methods, we propose a novel neural network framework, named Multi-View Scale Aggregation Networks (MVSAN), which comprehensively handle the scale variation from feature view, input view and criterion view. **In feature view**, we design a simple but effective Multi-Scale Feature Encoder (MSFE) to learn the scale robust representation. Specifically, our MSFE consists of three columns of CNNs, each of which

Corresponding author is Liang Lin.

This work was supported in part by the National Key Research and Development Program of China under Grant No.2018YFC0830103, in part by the National Science Foundation of China under Grant No.U1811463, No.61602533 and No.61702565, in part by the Fundamental Research Funds for the Central Universities under Grant No.18lgpy63.

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00259

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on January 06,2021 at 18:14:47 UTC from IEEE Xplore. Restrictions apply.

has four dilated convolutional layers. With different dilation rates, these columns have various receptive fields and can respectively model the appearance of people on different scales, thus boost the scale diversity of features. Moreover, with stacked MSFE, more feature from diversified scales are aggregated. **In input view**, we feed multiple scaled versions of an input image into our MVSAN to generate high-quality density maps in a coarse-to-fine manner. Our MVSAN is built upon two stacked CNNs, each of which is composed of a front-end Fully Convolutional Network (FCN) and three stacked Multi-Scale Feature Encoders. The first CNN takes the original image as the input and generates a coarse density map. Taking the image with high resolution as input, the second CNN refines the result of the first CNN and produces an accurate density map. **In criterion view**, we utilize a Multi-Scale Structural Similarity (MS-SSIM) loss to enforce our networks to learn the local correlation of multi-scale patches on the density maps. It is adapted from the Multi-Scale Structural Similarity Metric [4] and strengthened with dilated convolutional operations. Firstly, we build a criterion network with several fixed Gaussian kernel convolutional layers. Then, we feed the estimated density map and the ground truth map into the criterion network, and enforce SSIM loss between their output maps at every convolutional layer. To summarize, the main contributions of this work are three-fold:

- We propose a Multi-View Scale Aggregation Networks to comprehensively handle the huge variation of people scale. It integrates two front-end FCN and stacked Multi-Scale Feature Encoders to extract the scale robust feature and generate the density map in a coarse-to-fine manner.
- We design a novel training criterion, named Multi-Scale Structural Similarity (MS-SSIM) loss, which forces the networks to learn the local correlation of multi-scale patches on the density maps and generate high-quality density map and accurate crowd count.
- Extensive experiments and evaluations on two challenging benchmarks show that our proposed method achieves superior performance in comparison to other state-of-the-art methods.

2. RELATED WORKS

Recently, deep neural network has been widely used in urban management [5, 6, 7] and lots of inspiring crowd counting methods have been proposed. In this section, we provide a brief review of these deep learning based crowd counting methods. We will discuss these methods from aspects of the feature, input and criterion view.

1) Network design for feature extraction: Most of the current density map based methods endeavored to surmount the main obstacle of the large diversity of crowd scale and designed networks to extract multi-scale feature. Pioneeringly, Zhang et al. [1] proposed a multi-column convolution neural network with different receptive fields in different columns to learn the feature of different scales. Sam et al. [8] adopted the same multi-column network with an additional VGG network [9] which is trained to route input images to appropriate columns of the network to boost the multi-scale feature learning. Criticizing the limited scale adaptability of multi-column structure, Shen et al. [10] imitated a multi-scale U-Net architecture and Li et al. [3] aggregated multi-scale feature with dilated convolutions from the VGG feature. Lately, Deepak et al. [11] invented a growing network which could iteratively expand its model capacity to deal with diversified people scale. With these observations, we conclude that diversified multi-scale feature learning is of the highest significance on the estimation of crowd count and density map.

2) Integration of multiple input: Multi-scale inputs are a practical method widely applied in the area of computer vision including crowd counting. Onoro-Rubio et al. [12] proposed to use a pyramid of image patches of multiple scales to estimate the final density. With learning to rank between the count number of child-patches and parent-patches inputs, Liu et al. [13] integrated unlabeled crowd images to train their network for crowd counting.

3) Criterion design on crowd counting: Previous works considered less of the correlation between pixels on the density maps and dominantly applied Euclidean loss function to learn the pixel-wise regressions independently[1] [2] [3]. However, as criticized in [10], the Euclidean loss misled the network to generate blurring density maps with attenuated multi-scale feature encoding. Thus, a complicated compound of adversarial loss [14], perceptual loss [15], and Euclidean loss were used in [10] to overcome the unpleasant density map quality. But the generated density maps were still far from satisfactory. While Cao et al. [16] explored a combination of local structural consistency and Euclidean loss and obtained a state-of-the-art performance. But these methods penalized either the pixel-wise error or the mismatch of single-scale local structure instead of multi-scale local structural inconsistency, which is greatly expedient to the learning of density map generation for scale-diversified crowd scenes.

3. PROPOSED METHOD

3.1. Feature Aggregation with Multi-Scale Feature Encoder

In this section, we develop a unified neural network module, termed as Multi-Scale Feature Encoder (MSFE), to model the scale variation of people. Inspired by previous works [1, 17], we develop our MSFE with multiple columns of CNNs and each column is designed to handle the different range of scale variation. Differing from these methods, we implement our MSFE by dilated convolution layers [18] with different dilated rates instead of normal convolutional layers with different kernel sizes and channels in each column. As a good alternative of pooling layer, dilated convolutional layers use

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on January 06,2021 at 18:14:47 UTC from IEEE Xplore. Restrictions apply.

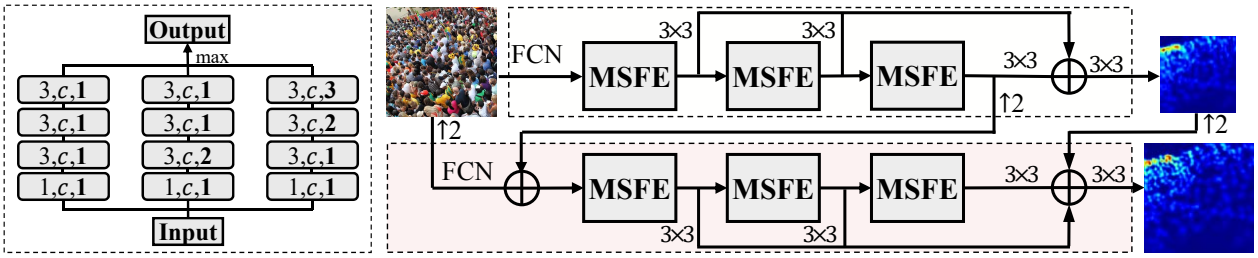

Fig. 2: **Left:** The architecture of the Multi-Scale Feature Encoder (MSFE). The block with text k, c, r denotes a dilated convolutional layer with $k \times k$ kernel size and c output channels. The dilation rate is expressed as r. **Right:** The architecture of the proposed multi-scale aggregation network with a coarse-to-fine scheme. "$\bigoplus$" denotes feature concatenation and "$\uparrow$ 2" refers to upsample by two times in width and height."3×3" refers to the convolutional layer to regress comprehensive density maps as described in Sec 3.1

sparse kernels to enlarge the receptive field without increasing the number of parameters or reducing the spatial resolution. In our MSFE, each column has different receptive fields by adopting particular dilation rates, yielding improved the scale diversity of features. The detail of MSFE is described as following. Similar to MCNN [1], our MSFE is composed of three columns of CNNs, each of which consists of four dilated convolutional layers. A ReLU layer is applied after every dilated convolutional layer. As shown in the left of Figure 2, the dilated convolutional layers at the same level have the same kernel size and channel number, and their parameters are shared. To learn the feature corresponding to heads of various scales, we apply different dilation rates in each column. In the first column, the dilated convolutional layers with dilation rate 1, which turns into normal convolutions, are used to extract the feature. The second and third columns utilize the convolutional layers with bigger dilation rates to enlarge the receptive field. Finally, we fuse the output features of these columns with an element-wise maximization operation to generate the scale robust representation.

3.2. Progressive Refinement with Multi-Scale Inputs

As shown in Figure 4, there may exist a mass of tiny heads in some public scenes. Most of the current methods utilize the convolutional neural networks with multiple pooling layers to generate density maps with low resolution, which are too rough to localize these tiny heads. To handle this issue, we design our network architecture with a coarse-to-fine scheme, which takes multiple scaled versions of the input image to generate a fine density map in high resolution via progressive refinement. In particular, our model is built upon two stacked CNNs. The first CNN takes the original image as the input to extract the feature and generate a coarse density map, roughly identifying crowd regions. Then, the image is resized to two times larger and fed into the second CNN to refine the feature of the first CNN, yielding improved the quality of density maps.

The detail of our network architecture is shown in the right of Figure 2. The first CNN consists of a front-end Fully Convolutional Network (FCN) and three stacked Multi-Scale Feature Encoder (MSFE). Specifically, the FCN is the first ten layers of VGG16 [9] with three pooling layers, while the channel number of each MSFE is set to 256. Given an image $I \in R^{H \times W}$, where H and W are the height and width of the image, we feed it into the network to extract the feature. We denoted the feature generated by i^{th} MSFE of the first CNN as f_1^i. Inspired by the DSN [19], for each MSFE, we feed its feature f_1^i into a convolutional layer with kernel size 3×3 to regress a density map m_1^i, which will help to accelerate the convergence with the deep supervision on side responses. Finally, we obtain a comprehensive density map $m_1 \in R^{\frac{H}{8} \times \frac{W}{8}}$ by feeding the concatenation of m_1^1, m_1^2 and m_1^3 into a weighted-fusion convolutional layer with kernel size 3×3. As shown in Figure 4, this density map m_1 can roughly localize crowd regions, but it fails to estimate the accurate number of people in the image. Thus, we utilize the second CNN to generate a fine density map.

The second CNN shares the similar network architecture of the first CNN, but it has much fewer parameters in the MSFE. Specifically, the channel numbers of there MSFE are 256, 128 and 64 respectively. Firstly, the image I is resized to $2H \times 2W$ via bilinear interpolation and fed into the front-end FCN of the second CNN, the output feature of which is denoted as $f_2^v \in \frac{H}{4} \times \frac{W}{4}$. Then, this feature is used to refine the output of the first CNN. After upsampled by two times in width and height, the feature f_1^3 is concatenated with f_2^v and fed into the three MSFE in the second CNN. Same with the first CNN, a side density map f_2^i is also computed after the i^{th} MSFE. Further, we concatenate f_2^1, f_2^2, f_2^3 and the upsampled density map $f_1 \in R^{\frac{H}{4} \times \frac{W}{4}}$ to produce the final density map f_2 with a 3×3 convolutional layer. As shown in Figure 4, f_2 can exhibit the accurate spatial location of the crowd. The whole coarse-to-fine scheme can infer in an end-to-end manner.

3.3. Local Correlation Learning with Multi-Scale Structural Similarity

Most of the previous approaches optimized their models with the pixel-wise Euclidean loss, which would cause the blurring effect in the density maps. Recently, Cao et al. [16] used a combination of Euclidean loss and single-scale structural similarity loss to train their networks, but their estimated maps still far from satisfactory. In this section, we utilize a Multi-

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on January 06,2021 at 18:14:47 UTC from IEEE Xplore. Restrictions apply.

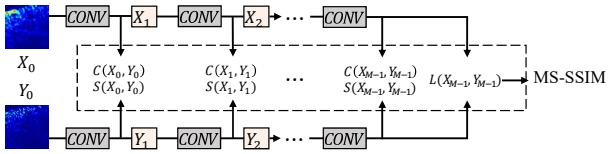

Fig. 3: Multi-scale structural similarity measurement. X_0 is the estimated density map and Y_0 is the corresponding ground truth. $CONV$ is the dilated convolutional layer with a normalized Gaussian kernel.

Scale Structural Similarity (MS-SSIM) loss to enforce our networks to learn the local correlation of multi-scale patches on the density maps. For convenience in the following, we denoted the estimated density map and the corresponding ground truth as X and Y respectively.

Single-Scale SSIM: As a common evaluation metric in image quality assessment, SSIM computes the similarity between two images/maps from three local statistics, i.e. mean, variance and covariance. Following [20], a 5×5 normalized Gaussian kernel with a standard deviation of 1.5 is used to estimate these local statistics. The estimation is easily implemented with a dilated convolutional layer with parameter $W = \{W(o)|o \in O, O_i = \{-2r, -r, 0, r, 2r\}, O_j = \{-2r, -r, 0, r, 2r\}\}$, where o is offset from the center and the dilation rate r is used to control the size of receptive field region in the Multi-Scale SSIM. For each location $p = (i, j)$ on X, the local mean $\mu_X(p)$, variance $\sigma_X^2(p)$ and covariance $\sigma_{XY}^2(p)$ can be computed by:

$$\mu_X(p) = \sum_{o \in O} W(o) \cdot X(p+o), \quad \sigma_X^2(p) = \sum_{o \in O} W(o) \cdot [X(p+o) - \mu_X(p)]^2, \quad (1)$$

$$\sigma_{XY}^2(p) = \sum_{o \in O} W(o) \cdot [X(p+o) - \mu_X(p)] \cdot [Y(p+o) - \mu_X(p)] \quad (2)$$

The mean $\mu_Y(p)$ and variance $\sigma_Y^2(p)$ of Y can be obtained with the same formulation. Further, the luminance comparison L, contrast comparison C and structure comparison S between X and Y can be calculated point by point as follows:

$$L(X,Y) = \frac{2\mu_X\mu_Y + c_1}{\mu_X^2 + \mu_Y^2 + c_1}, \; C(X,Y) = \frac{2\sigma_X\sigma_Y + c_2}{\sigma_X^2 + \sigma_Y^2 + c_2}, \; S(X,Y) = \frac{\sigma_{XY} + c_3}{\sigma_X\sigma_Y + c_3}; \quad (3)$$

where c_1, c_2 and c_3 are small constants to avoid division by zero.

Multi-Scale SSIM: As suggested in previous work [4], we argue that the estimated density map that is consistent with the ground-truth density map in multi-scale local correlation, better captures the crowd density distribution than those from solely pixel-wise or single scale consistency pursuits. With this insight, we build a CNN with M dilated convolutional layers, whose parameters are set to the fixed Gaussian kernel W described above. Specifically, M is set to 5 in our work and the dilation rates of these layers are 1,2,3,6 and 9 respectively. These layers are designed to calculate the SSIM of larger regions, which measure the local correlation comparison on multiple scales. As shown in Figure 3, after feeding X and Y into the CNN, we calculate the contrast and structure comparison after each layer. As in [4], the luminance comparison is computed only at the last layer. Finally, the MS-SSIM loss is defined as following:

$$MS\text{-}SSIM(X,Y) = [L_{M-1}(X_{M-1}, Y_{M-1})^{\alpha_{M-1}}] \cdot \prod_{j=0}^{M-1} [C_j(X_j, Y_j)^{\beta_j}] \cdot [S_j(X_j, Y_j)^{\gamma_j}], \quad (4)$$

$$MS\text{-}SSIM\ Loss = 1 - MS\text{-}SSIM(X,Y), \quad (5)$$

where the exponents α_j, β_j and γ_j are used to adjust the relative importance of different components and they are set with the same values as in [4]. Specifically, X_0=X and Y_0=Y as shown in Figure3. In our work, both the first and second CNN in proposed MVSAN are optimized with the MS-SSIM loss.

3.4. Implementation Details

Ground Truth Generation: During the training phase, we generate the ground truth density map with the geometry-adaptive kernels [1]. For each head annotation in a given image, assuming that the distances to its n nearest neighbors are denoted as $\{d_1, d_2, ..., d_n\}$, we label this head as a normalized Gaussian kernel with spread $\sigma = s \cdot \frac{1}{N} \sum_{i=1}^{N} d_i$. In our work, N is equal to 3 and the ratio s is set to 0.3. The radius of the Gaussian kernel is $6\sigma \times 6\sigma$. Some ground truth density maps are visualized in Figure 4.

Networks Optimization: We implement our crowd counting network with the Pytorch [21] toolbox. The two front-end FCN are initialized with the VGG [9] model pre-trained on ImageNet [22]. The filter weights of other convolutional layers are randomly initialized by Gaussian distributions with zero mean and standard deviation of 0.01. The learning rate is set to 1e-5. At each iteration, 16 patches with size 224×224 cropped from the images are fed into the networks. We first train the first CNN for 450 epochs and then jointly train whole networks for 250 epochs with Adam [23] optimizer.

4. EXPERIMENTS

4.1. Evaluation Metric

We evaluate the accuracy of crowd counting estimation with the widely adopted mean absolute error (MAE) and mean squared error (MSE), which are defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left|\hat{C}_i - C_i\right|; MSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left|\hat{C}_i - C_i\right|^2} \quad (6)$$

where N is the number of test samples, $\hat{C}_i$ is the estimated crowd count as the sum of all pixel values on the generated density map and C_i is the ground true crowd count. Moreover, the quality of density maps is measured with two standard metrics: SSIM and PSNR [24].

4.2. Evaluations and Comparisons

In this section, we evaluate and compare our proposed method with other state-of-the-art methods in the following representative datasets of relatively sparse, congested and highly congested crowd scenes.

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on January 06,2021 at 18:14:47 UTC from IEEE Xplore. Restrictions apply.

Table 1: Model Evaluation

(a) ShanghaiTech

Method	Part A		Part B	
	MAE	MSE	MAE	MSE
MCNN [1]	110.2	173.2	32.0	49.8
Cascaded-MTL [25]	101.3	152.4	20.0	31.1
Switch-CNN [8]	90.4	135.0	21.6	33.4
CP-CNN [2]	73.6	106.4	20.1	30.1
CSRNet [3]	68.2	115.0	10.6	16.0
SANet [16]	67.0	104.5	8.4	13.6
Ours	**59.2**	**87.2**	**9.0**	**14.0**

(b) UCF-QNRF

Method	MAE	MSE
Idees et al. [26]	315	508
MCNN [1]	277	426
Cascaded-MTL [25]	252	514
SwitchCNN [8]	228	445
Resnet101 [27]	190	277
Densenet201 [28]	163	226
Idees et al. [29]	132	191
Ours	**102**	**170**

Table 2: Quality of Density Maps

Method	SSIM	PSNR
MCNN [1]	0.52	21.4
CP-CNN [2]	0.72	21.72
CSRNet [3]	0.76	23.79
Ours	**0.77**	**23.17**

ShanghaiTech: ShanghaiTech crowd counting dataset [1] is composed of Part_A and Part_B. Part_A contains 482 congested images crawled from the Internet with an average amount of about 500 annotations every image. Part_B collects 716 images of relative sparse crowd scenes from streets containing roughly 123 annotated peoples on average. The comparison to other six recent works is shown in Table 1a. Our method achieves the lowest MAE and MSE in Part_A with 11.6% and 16.5% lower error than the existing best method respectively. In Part_B, our method outperforms most methods and achieves a comparable performance with best-performing method SANet [16], which uses patch-based inference. The outstanding performance indicates the accuracy and robustness of our method on the crowd counting on both congested scenes and relatively sparse scenes. Meanwhile, as shown in Table 2, the generated density maps of our method has the highest average SSIM than the other three methods and a close PSNR with the current best-performing method. The generated density maps of our method yield high image quality as well as accurate count estimations.

UCF-QNRF: The lately proposed UCF-QNRF dataset [29] contains 1535 images of highly congestive scenes and about 815 annotations per image on average. This dataset contains the most diverse set of viewpoints, densities and people scales and becomes the most challenging dataset. Following [29], we compare our method with other seven state-of-art methods. As shown in Table 1b our method achieves the lowest error metrics and outperforms current best-performing methods with 22.7% lower MAE and 10.9% lower MSE. To summarise, our proposed MVSAN can accurately count the number of crowds from relative sparse, congested to hight congested scenes better than current state-of-the-art methods.

4.3. Ablation Study on ShanghaiTech Part A

Feature view multi-scale aggregation: Three networks, directly regression network from the front-end FCN feature and the proposed coarse network with MSFE in one column (the left column on Figure 2) and three columns, are constructed and trained under the proposed MS-SSIM loss. Their performances in the Part_A of ShanghaiTech are compared in Table 3b. The network that directly regresses the density map from the front-end FCN feature performs poorer than the other two with MSFE. With more scale variance receptiveness than one-column MSFE, the proposed three-column MSFE achieves the lowest MAE and MSE. Specifically, the proposed three-column MSFE achieves 6.6% and 9.6% lower MAE and MSE than the network without MSFE.

Table 3: Ablation Study

(a) Different Criterions

Scale	Criterion	MAE	MSE
Single	L2	68.9	112.7
	SSIM[20]	79.8	140.3
	SSIM + L2[16]	68.3	109.8
Multi	MSSSIM-O[4]	65.8	110.1
	MSSSIM-D	61.8	96.6

(b) Effectiveness of MSFE block

Network	MAE	MSE
W/O MSFE	66.2	107.0
W/ Single-Column MSFE	64.5	100.1
W/ Three-Column MSFE	61.8	96.6

(c) Multi-scale Inputs Validation

Input Scale	MAE	MSE	SSIM	PSNR
Single $\times 1$	61.8	96.6	0.72	21.39
Single $\times 2$	66.8	111.3	0.78	24.03
Multi $\times 1$,$\times 2$	59.2	87.4	0.77	23.17

Criterion view multi-scale aggregation: Best performance of the coarse network trained with different criteria of single scale and multiple scales are listed in Table 3a. Single scale criteria include Euclidean criterion(L2), SSIM and their combination as in [16]. Meanwhile multi-scale criteria include the original MS-SSIM(MSSIM-O) as in [4] and our proposed MS-SSIM(MSSSIM-D), which innovatively replaces the low-pass filter and downsample operations on the original one with dilation convolution. Performance of model trained under single scale criteria including is not comparable with those trained under Multi-scale criteria. Specifically, our designed MSSSIM-D achieves 7.2% lower MAE and 13.9% lower MSE than the original MSSSIM-O baseline.

Input view multi-scale aggregation: We compare the performance of our network trained in the coarse-to-fine scheme with multi-scale inputs and its two sub-networks trained with single scale input, the original images and the upsampled images respectively. The sub-network operating in original images is termed as coarse sub-network and the other is denoted as refiner sub-network in the following. As presented in Table 3c, the coarse sub-network and the refiner sub-network perform poorer than the whole MVSAN trained with multi-scale inputs. With the trained coarse sub-network and additional input scale to the refiner network, our network trained in a coarse-to-fine scheme achieves 4.2% and 9.5% lower MAE and MSE than the coarse network. The image quality is also greatly improved. Three testing samples and their ground truth density maps, and the estimated density maps of the coarse sub-network and the whole network are presented in Figure 4. Limited by single scale input, the coarse sub-network is unable to distinguish people in congested place separately and responses with the continuous high-response burring areas along with inexact counts. These areas are refined with the aid of the refiner network using additional input scale as shown in Figure 4. We could observe the progressive refinement on density maps and count estimations with multi-scale inputs from left to right in Figure 4.

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on January 06,2021 at 18:14:47 UTC from IEEE Xplore. Restrictions apply.

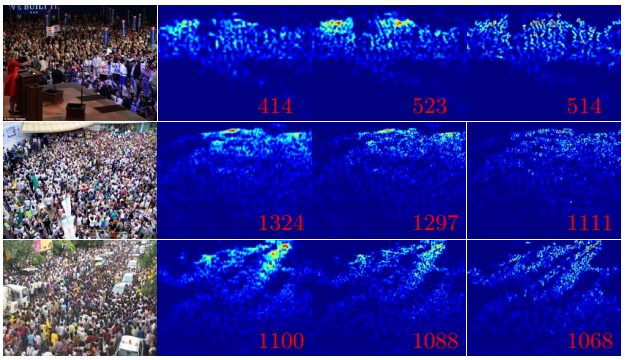

Fig. 4: Original crowd scenes, estimations of coarse network and final estimations and the ground truth density maps are displayed from left to right. People counts are labeled in red on images. **Zoom in on details.**

5. CONCLUSION

In this work, we propose a novel multi-view scale aggregation network(MSVAN) for high-quality density maps generation and accurate crowd count estimation. It comprehensively handles the scale variation from feature view, input view and criterion view. In the feature view, the proposed multi-scale feature encoder(MSFE) effectively encodes scale-diversified feature and integrates to front-end convolutional feature extractor, which together make up the sub-networks of our MSVAN. In the input view, with scaled versions of an image as inputs, sub-networks aggregate features and coarse density maps in lower resolution from their preceding sub-network to generate refined density maps. In the criteria view, we design a multi-scale structural similarity criterion to enforce our MSVAN to exploit local correlation on the patches of varying scales to generate high-quality density map. Experiments on two datasets of crowd scenes with various crowd density show the superior performance of our methods over the state-of-the-art methods with accurate estimations and high-quality density maps.

6. REFERENCES

[1] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma, "Single-image crowd counting via multi-column convolutional neural network," in *CVPR*, 2016.

[2] Vishwanath A Sindagi and Vishal M Patel, "Generating high-quality crowd density maps using contextual pyramid cnns," in *ICCV*, 2017.

[3] Yuhong Li, Xiaofan Zhang, and Deming Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *CVPR*, 2018.

[4] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *ACSSC*, 2003.

[5] Shanghang Zhang, Guanhang Wu, Joao P Costeira, and José MF Moura, "Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras," in *CVPR*, 2017.

[6] Zhilin Qiu, Lingbo Liu, Guanbin Li, Qing Wang, Nong Xiao, and Liang Lin, "Taxi origin-destination demand prediction with contextualized spatial-temporal network," in *ICME*, 2019.

[7] Lingbo Liu, Ruimao Zhang, Jiefeng Peng, Guanbin Li, Bowen Du, and Liang Lin, "Attentive crowd flow machines," in *ACM MM*, 2018.

[8] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu, "Switching convolutional neural network for crowd counting," in *CVPR*, 2017.

[9] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[10] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang, "Crowd counting via adversarial cross-scale consistency pursuit," in *CVPR*, 2018.

[11] Deepak Babu Sam, Neeraj N Sajjan, R Venkatesh Babu, and Mukundhan Srinivasan, "Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn," in *CVPR*, 2018.

[12] Daniel Onoro-Rubio and Roberto J López-Sastre, "Towards perspective-free object counting with deep learning," in *ECCV*, 2016.

[13] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov, "Leveraging unlabeled data for crowd counting by learning to rank," *arXiv preprint arXiv:1803.03095*, 2018.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.

[16] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su, "Scale aggregation network for accurate and efficient crowd counting," in *ECCV*, 2018.

[17] Lingbo Liu, Hongjun Wang, Guanbin Li, Wanli Ouyang, and Liang Lin, "Crowd counting using deep recurrent spatial-aware network," in *IJCAI*, 2018.

[18] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[19] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu, "Deeply-supervised nets," in *Artificial Intelligence and Statistics*, 2015.

[20] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *TIP*, 2004.

[21] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," 2017.

[22] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[23] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] Quan Huynh-Thu and Mohammed Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, 2008.

[25] Vishwanath A Sindagi and Vishal M Patel, "Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *AVSS*, 2017.

[26] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *CVPR*, 2013.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[28] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks.," in *CVPR*, 2017.

[29] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah, "Composition loss for counting, density map estimation and localization in dense crowds," *arXiv preprint arXiv:1808.01050*, 2018.

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on January 06,2021 at 18:14:47 UTC from IEEE Xplore. Restrictions apply.