Multi Scale Attention Network for Crowd Counting

Xiangpeng Yang
School of Automation, Southeast University, Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, China
knightyxp@gmail.com

Xiaobo Lu
School of Automation, Southeast University, Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, China
xblu2013@126.com

ABSTRACT

Reasonable management and control of extra crowded scenes have become a hot topic in recent years. Counting people from density map generated from the object location annotations is an effective way to analyze crowd information and control crowds in severely congested scenes. In this paper, we propose a novel end-to-end crowd counting method called MSANet for crowd counting. MSANet consists of the VGG16 backbone as the fronted part, two branches as the back-end part, including the attention map extractor to predict crowd states (means with people or not), and density map branch to regress the density map. What is more, to obtain high-resolution density map, we combine different scale maps from the front part to the back-end part. On the design of the loss function, to enhance the resolution of the predicted map and its structural similarity to ground truth, we proposed a new loss function for crowd counting. The test result based on the public dataset ShanghaiTech and Subway Crowd Counting Dataset supported by the Nanjing Metro demonstrates the effectiveness of our method.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**; • **Computer vision problems**;

KEYWORDS

Crowd counting, Multi scale attention network, Density map, Attention map, Structural similarity

ACM Reference Format:
Xiangpeng Yang and Xiaobo Lu. 2021. Multi Scale Attention Network for Crowd Counting. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3487075.3487097

With the development of social economy, large-scale crowded scenes have become more common. Crowd counting is an effective way for crowd understanding and analysis, public safety and traffic control. However, due to the occlusion caused by viewing angle, perspective distortions in public surveillance cameras and scale inconsistency of pedestrians, crowd counting has become a challenging problem in the computer vision field.

In terms of crowd counting technology, early methods take advantage of hand-crafted features including Harr wavelets [1], histogram oriented gradients [2] and so on to tackle the modeling of pedestrians. Inspired by deep learning and convolutional neural network, some classic detection-based approaches locating each people by detecting each person's face, like tiny face detector [3] and SSH [4]. However, severe occlusion and multi-scale distribution of people in crowded scene make these methods could only work in the foreground area of an image, hard to detect extremely tiny heads whose sizes are about few pixels.

Till now, the mainstream of crowd counting methods is based on the regression of density map. Lempitsky [5] firstly developed to treat crowd counting problem as a density map regression problem, the density map is generated by filtering the dot-level annotation map through normalized 2D gaussian kernel, the sum of each pixel value is the count of the whole map. To deal with the different sizes of head, MCNN [6] utilizes three-stream convolution network with different convolutional kernel sizes. CSRNet [7] exploited to single-column model to reduce parameters and creatively applied dilated convolution to increase the receptive field of the counting network. CANet [8] innovatively proposes a context-aware module which means using contrast features obtained by the pooling pyramid to assign different weight attention of the density map.

However, in actual crowded scenes such as subway entrance or large crowd gatherings in ShanghaiTech dataset [6], due to the occlusion of high-density regions, Undifferentiated regression of the density of each pixel is unreasonable. Due to the current annotation scheme, the values in the density map have two states: zero means no objects existed in the pixel's neighborhood, while a non-zero value indicates the local density. From the motivation of removing background noise, we devise a separate branch to regress the attention map with the supervision of cross entropy loss. This action brings two benefits: 1. the density estimation, a regression problem, could be better trained with the aid of segmentation loss. 2. Attention map making the network focus on the area with crowds, which is obvious better than treat each field in the density map equally. Moreover, to ensure the network's ability to extract multi scale features, we exploit the FPN [9] structure and adopt a multi scale structural similarity loss function to ensure the local coherence of density map.

Nowadays, subway station has become the most import transportation tools for urban people. How to count the number of pedestrians accurately in the subway surveillance videos to prevent crowd gathering or trampling incidents, and to ensure the normal states of subway traffic is a very essential topic. However, existing

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487097

public datasets such as ShanghaiTech dataset, UCF-QNRF [10] do not contains any scenes in subway, streets are the most common scenes in current public datasets. Therefore, we propose a new crowd counting dataset name Subway Crowd Counting Dataset (SCCD) with 5 common subway scenes in the Nanjing subway station to meet the needs of subway pedestrian counting.

To sum up, our contribution can be summarized as follows:

- From the view of enhancing the network's ability to extract multi scale effective information, we exploit dilated convolution in encoder block and combine FPN structure with attention map.
- To keep the local coherence of the predicted density map with ground truth, enlightened by CFANet [11], we designed a Average Multi-Scale Structural Similarity loss (Avg MS-SSIM loss) for crowd counting.
- We construct a new dataset called Subway Crowd Counting Dataset (SCCD) which is supported by Nanjing Metro. The proposed MSANet performs better than the state of art on SCCD and public crowd counting dataset, which verified our MASNet is more robust and accurate for crowd counting.

1 RELATED WORK

Recently, there are plenty of works about crowd counting. Here we only summarize the works which are research on similar problems: multi scale counting network, multi learning objectives counting, structural similarity in crowd counting and subway crowd counting

Multi Scale counting network: PACNN [12] proposes multi scale perspective-aware network for crowd counting, which is inspired by the perspective geometry of a pinhole camera. MBTTBF [13] mainly explore how to efficiently leverage different scale information, they present a multi-level bottom-top and top-bottom fusion scheme to effectively merge information from multiple network layers. Multi column Mutual Learning [14] proposes a mutual learning scheme named McML to update parameters of multiple columns network asynchronously, which is a supplement to the inadequate extraction of multi-scale information from the multi-column structure. MSPNET [15] is consisted of VGGG backbone and a dual-stream decoder network with 3 different sizes of ground truth to conduct multiple supervisions. ASD [16] contains 3 branches: sparse branch, dense branch and adaption branch. The high light of this work is applying excitation operation in SENet [17] to crowd counting, a series operation including global average pooling and fully connected layers to gain the global crowd information. bi-path optical flow-based crowd count network [18] is a network design for cross-scene stability and adopts a mass of synthetic data in remote crowd count tasks.

Multi learning objectives counting: W-net [19] apply an encoder-decoder network similar to U-net [20] in crowd counting, they train the Reinforcement branch as a classification task to make the whole network coverage faster. CFF [21] adopt a 3-branches network with 3 tasks including segmentation map regress with BCE loss, global density map with focal loss [22], and density map with MSE loss for density estimation. ADCrowdNet [23] uses a two-step model for crowd counting, attention map is generated in the AMG model, DME model inputs combine the input image and attention map by pixelwise product, then go through an encode-decoder network to produce final density map.

Structural similarity in crowd counting: SANet [24] firstly adopt local pattern consistency loss which is a kind of structural similarity loss computation in each fix-sized gaussian kernel in crowd counting deep learning methods. DSSINet [25] exploited to refine U-net and creatively adopt the dilation structural similarity and conditional random field to refine the fusion method of multi scale crowd deep feature. SE Cycle GAN [26] use the original SSIM loss for remedying the problem that origin cycle consistency is more easily to lose local patterns and text features in the task of recalling crowd counting via domain adaptation.

Subway crowd counting: DRL (double-region learning) [27] proposes the Subway Station pedestrian dataset (SSPD), SSPD Only includeS one scene condition of transfer corridors in 5 different stations including Jinshajiang Road, Jing'an. Temple, Shanghai South Railway Station, People Square and Xujiahui.

In this paper, we propose a multi scale attention network for crowd counting. In terms of network design, our MSANet is somehow similar to W-net. However, we adopt dilated convolution in decoder network to enlarge receptive field furthermore. To clarify, our method extracts multi scale effective information not only by FPN structure and attention map, but also through the Avg MS-SSIM loss to pooling the density map several times into different sizes and compute the local structural similarity by sliding window to keep local consistency between predicted map and ground truth.

2 OUR PROPOSED METHOD

The overview of MSANet is illustrated in the Figure 1. Our MSANet could be divided into two parts, the front part is called Feature Map Extractor, the end part is a dual-stream network, consisted of Attention Map Extractor (AME) and Density Map Extractor (DME). In terms of the network's architecture design, the top-to-down feedback manner adopted in MSANet is similar to Mask-Aware-Network [28] and other methods using attention map [29], [11].

Different from above methods, the highlights of our MSANet can be summarized as following three points. Firstly, the back-end part is combined with dilated convolution layers to capture more accurate high-level features with larger receptive. Furthermore, the back-end part design makes two extractors is sharing parameters, which automatically forcing the density map to focus on the foreground information. Additionally, FPN structure and Avg MS-SSIM loss make our model adaptively aware multi scale density information.

2.1 Feature Map Extractor (FME)

Following the paradigm of previous studies, the FME employs the first 13 Convolutional layers of conv1_1 to conv5_3 in a pre-trained VGG16-bn and 4 Maxpooling layers. These layers are divided into 4 parts as shown in Figure 1. Since the 4 times Maxpooling operation with filter size 2x2, the feature maps' size of P1, P2, P3, P4 is 1/2, 1/4, 1/8, 1/16 of original feature map.

2.2 Dilated Convolution and FPN

The architecture of back-end part is enlightened by the Feature Pyramid Network (FPN) and the dilation convolutional layer used in CSRNet.

To be more specified, the FPN structure leverage hierarchical pyramids which just is a top-down CNN with skip connections to enhance the network's ability to aware of multi-scale semantic features. Therefore, in our implementations, we concatenate same size of feature map in front part and back-end part.

As for the convolutional layer, to handle the problem that extremely dense areas' feature where one-pixel region can contain several people can not be extracted suitably. We apply the dilated convolution in back-end part to acquire more detailed information. A normal 2-D dilated convolution can be defined as:

$$output\,(m,n) = \sum_{i=1}^{M}\sum_{j=1}^{N} input\,(m + r \times i, n + r \times j)\, w\,(i,j) \qquad (1)$$

Where $w(i, j)$ is a convolution filter with the length of M and width of N respectively, r is the dilated rate.

Dilated convolution is more efficient, a filter of $k \times k$ kernel size with dilated rate r could be regarded as a filter with receptive filed of $K + (K-1)(r-1)$. Consequently, we replace the 3x3 normal convolutional layers in the back-end part with 3x3 dilated convolutional layers. In this way, the receptive field can be enlarged from 3x3 to 5x5. What is more, to maintain the resolution of output feature map unchanged, we try dilated convolution of kernel size of 3x3 with dilated rate =2 and padding=2.

2.3 Attention Map Extractor (DME)

Considering that constructions and obstacles in an image will force the crowd to be distributed in the open area, to make the network only compute the scene where has people and avoid the situation that blank areas are counted but the total count of an image still is right, Attention Map Extractor comes to the sights of ours.

The configuration of AME is U-C(1024,256,1)-D(256,256,3)-U-C(512,128,1)-D(128,128,3)-U-C(256,64,1)-D(64,64,3)-D(64,32,3). where C(x, y, 1) means normal convolutional layer with input channels of x and output channels of y, the kernel size is 1x1 and stride is 0, D(x, y, 3) is a dilated convolutional layer with input channels of x and output channels of y, the kernel size is 3x3 and padding is 2, U means bilinear upsampling with scale factor 2. Due to the 3 times upsampling operation, the estimated attention map is half size of the original input. The Attention map is generated by sigmoid activation to distinguish whether a pixel in the density map has crowds. The Structure from P1* to the predicted attention map is normal convolution layer of kernel size 1x1 and sigmoid activation.

2.4 Density Map Extractor (DME)

The DME's structure is same as attention map. Actually, the two extractors are sharing parameters by using the same structure. In detail, the way to combine the attention map and density map is element-wise multiplication and a convolutional layer of kernel size 1x1, which is the milestone experimental result of Mask Aware Network. Follow the paradigm of using attention map, this process

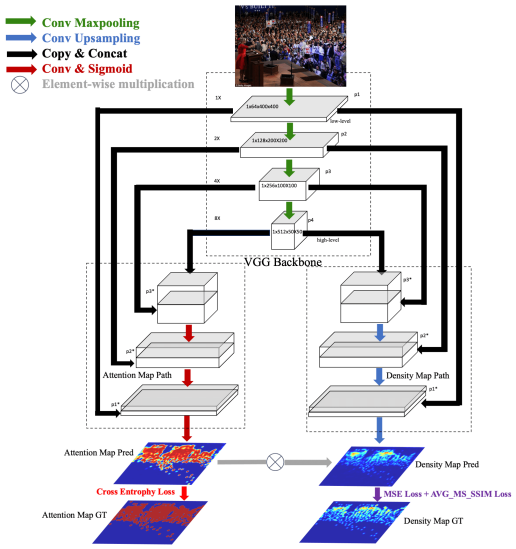

Figure 1: The Architecture of Proposed MSANet.

can be defined as:

$$DM_{refined} = DM_{output} \otimes AM_{output} \qquad (2)$$

Where DM_{output} is the output of DME, AM_{output} is the output of AME, $DM_{refined}$ is the final density map, $\otimes$ means elementwise multiplication.

Moreover, we adopt Batch Normalization [30] and ReLU in each convolutional stage to stabilize the training process and prevent overfitting.

2.5 The Loss Function

As shown in Figure 1, MSANet's loss function design are composed of 3 different loss.

To discriminate the crowd states and considering the fact that the attention map only has two values, 0 means no crowd in one pixel and 1 means have crowds. Therefore, in the AME, we treat the attention map prediction problem as a binary segmentation problem, the loss function of AME is binary cross entropy loss. The BCE loss can be defined as:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}(am_i \times \log(p_i) + p_i \times \log(am_i)) \qquad (3)$$

Where N is the total number of images in a batch, am_i is the label of the predicted attention map (0/1), p_i is the possibility the pixel is predicted to have crowds.

Also, to ensure high resolution of final density map and its similarity to the ground truth, we adopt the common mean square loss (MSE) which is widely used in crowd counting and regression task. The MSE loss is formulated as:

$$L_{MSE} = -\frac{1}{N}\sum_{i=1}^{N}(dm(x) - dt\,(x))^2 \qquad (4)$$

Where $dm(x)$ is the predicted fine-grained density map, $dt(x)$ is the ground truth of the density map.

Additionally, from the perspective of forcing the DME to study the local correlation of different density crowd regions with ground truth and to reduce the false recognition of predicted density map. Inspired by the local pattern consistency loss in SANet, DMS-SSIM loss in DSSINet and the BSL (Background-aware Structural Loss) in CFANet, we devise a list of experiments to find the most efficient structural similarity loss—Average Multi-Scale Structural Similarity loss function (Avg-MS-SSIM loss) in crowd counting.

To keep the local coherence of estimated density map while extracting multi scale information, Avg-MS-SSIM loss pooling the original density map into different sizes (1/2,1/4,1/8 etc) and then use an 11×11 normalized Gaussian kernel with a standard deviation of 1.5 to estimate local coherence. The local coherence is ensured by the original structural similarity loss computation in each 11x11 sliding window.

Our experiment in section 4.3.2 suggests that pooling predicted density map 3 times is the most powerful in Avg-MS-SSIM loss.

$$L_{Avg-MS-SSIM} = -\frac{1}{K}\sum_{j=1}^{k}\left(1 - L_{SSIM}\left(pool_j\left(dm\right) - pool_j\left(dt\right)\right)\right)^2 \tag{5}$$

$$L_{SSIM} = 1 - \frac{\left(2\mu_x\mu_{y+}c_1\right)\left(\sigma_{xy} + c_2\right)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)} \tag{6}$$

Where j represents the map size after pooling is $\frac{1}{2^{j-1}}$, the local means are μ_x, μ_y and the σ_x, σ_y represent local variance respectively, σ_{xy} is the local covariance. The hyperparameters setting is as followings: $c_1 = 0.01$, $c_2 = 0.03$. The most essential parameter is the average frequency K is set to 3, according to the ablation studies in section 4.3.2.

Finally, the loss function of MSANet is defined as:

$$L_{MSAN} = L_{MSE} + L_{Avg-MS-SSIM} + \lambda \times L_{BCE} \tag{7}$$

Where λ is the weight of BCE loss, set to 0.1.

3 EXPERIMENTS AND ANALYSIS

In this section, we will introduce the evaluation metrics and ground truth generation, 3 datasets including Shanghai Tech dataset, UCF-QNRF and Subway Crowd Counting dataset constructed by ourselves, our MSANet's performance on the 3 datasets, and ablation studies.

3.1 Ground Truth Generation and Evaluation

3.1.1 Ground Truth Generation. Ground truth generation is comprised of density map ground truth and attention map ground truth.

For the ground-truth density maps, follow the guidelines in MCNN, each head annotation $\delta(x - x_i)$ is replaced by a length fixed standard deviation Gaussian kernel $G_\sigma(x)$:

$$D_{gt} = \sum_{i=1}^{C}\delta\left(x - x_i\right) \times G_\sigma\left(x\right) \tag{8}$$

Where c represents the total numbers of head annotations and σ is an empirically chosen variance term. The choice of σ should make $G_\sigma(x) = 0$ if x is out of the range of local neighborhood x_i. In the 3 experiment datasets, σ is set to 5. Since the kernel is normalized with standard deviation, the integral of the whole density map is equal to the total number of people.

For the ground truth attention maps, they are generated based on the ground-truth density map. As a solution to discriminate background and foreground, if the counting number is larger than the threshold 1e -5, the value of attention map is set to be 1, else to 0, the formulation is defined as:

$$A_{gt}\left(x\right) = \begin{cases} 0 & D_{gt}\left(x\right) < 1e-5 \\ 1 & D_{gt}\left(x\right) > 1e-5 \end{cases} \tag{9}$$

Where x means the position coordinate in the ground-truth density map.

3.1.2 Evaluation Metrics. We adopt Mean Absolute Error (MAE) and Mean Square Error (MSE) for the counting accuracy metrics:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|pred_i - gt_i\right| \tag{10}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(pred_i - gt_i\right)^2} \tag{11}$$

Where N is total number of images in the test dataset, $pred_i$ is the predicted density map of network towards to the i-th image, while gt_i is the ground truth density map of the i-th image.

3.2 Datasets and Performance

We evaluate our proposed MSANet on the challenging public dataset ShanghaiTech, UCF-QNRF and practical scene, Subway Crowd Counting dataset.

3.2.1 ShanghaiTech. ShanghaiTech dataset contains 1198 images with 330165 annotated heads, which can be divided into 2 parts, ShanghaiPart A and Part B. Part A has 300 training images and 182 testing images randomly collected from internet and is more congested compared to the other part. Part B has 400 training images and 316 testing images collected from the shanghai streets with relatively sparse scenes.

From Table 1 and Table 2, it can be seen that our MSANet with Avg-MS-SSIM loss achieves the most fascinating results.

As shown in Table 1, MSANet's MAE on ShanghaiTech part A is enhanced by 1.02% compared to the state of the art. Meanwhile, MSANet achieves the third-best MAE on Part B sub-dataset.

The representative visualization results of MSANet on ShanghaiTech Part A and Part B are shown in Figure 2 and Figure 3. In the 2 figures, the first column is the sample images, the second column is the predicted density map, the last column is the ground truth of the density map

3.2.2 UCF-QNRF.. UCF-QNRF is the most largest crowd counting dataset, including 1535 images (1201 for train and 334 for test). The total head point annotations are 1.25 million, people count range from 49 to 12865. Since the average image's resolution in UCF-QNRF is as high as 2013x2902, we resize each image in train dataset to make sure the longer side falls into [1024,2048], while the images keep the same in the testing process.

As shown in Table3, Our MSANet gains the best performance when compared to the methods mentioned in this paper, surpassing

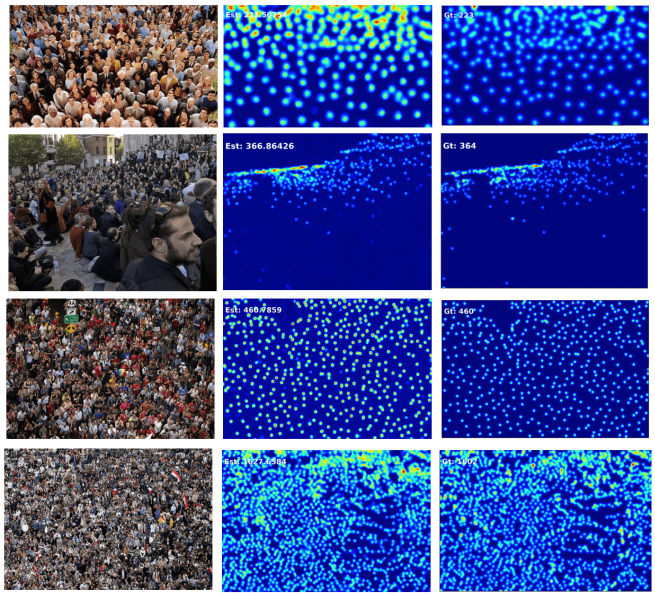

Figure 2: Visualization of Predicted Density Maps on ShanghaiTech Part A.

Table 1: Performance Comparison on ShanghaiTech Part A

Method	MAE	MSE
MCNN	110.2	173.2
CSRNet	68.2	115.0
SANet	67.0	104.5
ASD	65.6	98.0
CFF	65.2	109.4
TEDNet [31]	64.2	109.1
ADCrowdNet(AMG-bAttn-DME)	63.2	98.9
PACNN+	62.4	102.0
CANet	62.3	100.0
Bayesian loss	62.8	101.8
DSSINet	60.6	96.0
Mask-aware Net	61.8	100.0
MBTTBF-SCFB	60.2	94.1
MSPNet	59.8	98.2
W-Net	59.5	97.3
CSRNet+McML	59.1	104.3
Our MSANet	**58.5**	**98.5**

the second best approach by 2.1% in MAE. The visualization of representative results on UCF-QNRF is shown in Figure 4

3.2.3 Subway Crowd Counting Dataset. With the support of Nanjing Metro, Subway Crowd Couting Dataset (SCCD) is built. SCCD is based on the surveillance video of Nanjing Metro, we extract one frame as an image at an interval of 30 seconds.

To meet the requirement of subway crowds monitoring, as shown in Figure5, we selected 5 representative subway scenes that are most likely to have dangerous events: security check, subway platform, subway escalator, subway entrances/exit, transfer corridors in 3 different stations (Nanjing Station, Liuzhoudonglu Station, Youfangqiao Station). The SCCD including sparse, medium, and exceeding congested scenes in subway. The train and test datasets are composed of 6014 images and 1150 images respectively.

Compared with CSRNet, our MSANet achieves attractive performance on SCCD, the specific details are listed in Table 4

From Table 4, compared to CSRNet, the contrastive results of 6.7% MAE relative improvement on SCCD demonstrates the effectiveness of MSANet. The visualization of MSANet's predicted results on SCCD is shown in Figure 6

Table 2: Performance Comparison on ShanghaiTech Part B

Method	MAE	MSE
MCNN	26.4	41.3
CSRNet	10.6	16.0
SANet	8.4	13.6
TEDNet	8.2	12.8
CSRNet+McML	8.1	10.6
Mask-aware Net	8.6	13.3
ASD	8.5	13.7
ADCrowdNet(AMG-bAttn-DME)	8.2	15.7
MBTTBF-SCFB	8.0	15.5
CANet	7.8	12.2
Bayesian loss	7.7	12.7
PACNN+	7.6	11.8
MSPNet	7.5	14.1
CFF	7.2	12.2
DSSINet	6.9	10.3
W-Net	6.9	10.3
Our MSANet	**7.0**	**11.1**

Table 3: Performance Comparison on UCF-QNRF

Method	MAE	MSE
TEDNet	113.0	183.0
CANet	107.0	188.0
DSSINet	99.1	159.2
MBTTBF-SCFB	97.5	165.2
CFF	93.8	46.5
Bayesian loss	88.7	154.8
Our MSANet	**86.9**	**151.3**

3.3 Ablabtion Study

In this section, we implement a series of ablation studies find the most suitable output size of MSANet, how many times of pooling is best for Avg-MS-SSIM loss, to evaluate the effectiveness of Avg MS-SSIM loss.

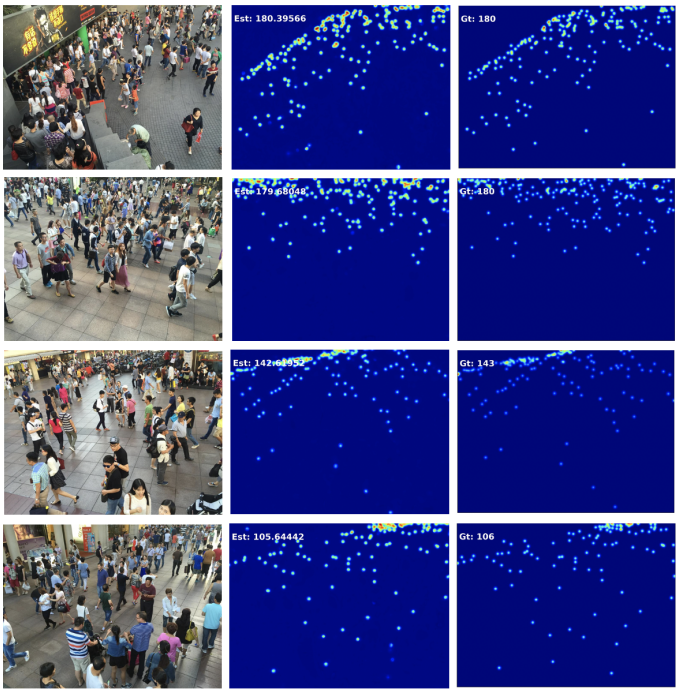

Figure 3: Visualization of Predicted Density Maps on ShanghaiTech Part B

Figure 4: Visualization of Predicted Density Maps on UCF-QNRF.

3.3.1 Output Size. Whether the output size is same as input or half size is better for density estimation is a problem. Therefore, we design a similar network with same size output as input, called MSANet org_szie, while MSANet with half-size output is the baseline. In the settings of the MSANet org_szie, feature map extractor (FME) keep same as baseline, the density map extractor is still same as attention map extractor with following settings: U-C(1024,256,1)-D(256,256,3)-U-C(512,128,1)-D(128,128,3)-U-C(256,64,1)-D(64,64,3)-U-C(128,32,1)- D(32,32,3)- D(32,16,3), other settings is unchanged.

Table 4: Performance Comparison on SCCD

Method	MAE	MSE
CSRNet	4.16	8.88
Our MSANet	**3.88**	**8.15**

Table 5: Ablation Study of Output Size

Network	MAE	MSE
MSANet org size	60.65	102.65
Baseline	**58.55**	**98.45**

Figure 5: Illustration of Subway Crowd Counting Dataset Scenes.

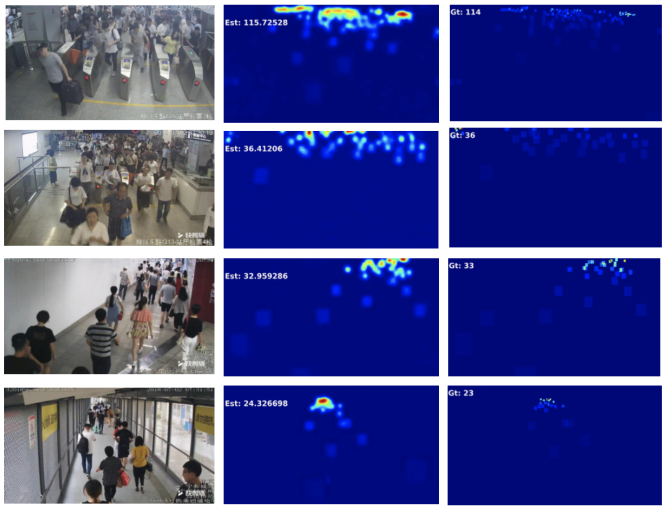

Figure 6: Visualization of Predicted Density Maps on SCCD.

As shown in Table 5, we compare the MSANet org_size's performance on ShanghaiTech Part A with baseline. According to the result of experiment, half-size output is better than original size output with 3.8% improvement on MAE.

3.3.2 Pooling Times. Pooling times is a crucial hyper-parameter in Avg MS-SSIM loss, too many pooling times will lead to excessive calculation overhead, and too few times will degenerate into the original multi-scale structure similarity loss.

From the Table 6, we can see that pooling 3 times could achieve the best performance on ShanghaiTech Part A.

3.3.3 The Effectiveness of Avg MS-SSIM Loss. Table 7 is MSANet's ablation study of loss function, baseline means without any kind of structural similarity loss, the second row means MASNet with original similarity loss, the third row represents MSANet with the Avg-MS-SSIM loss at the average level of 3. Compared to the baseline, the Avg-MS-SSIM loss gains 8.2% and 30% relative MAE improvement on part A and part B. When calculating different scale map's structural similarity, means using Avg-MS-SSIM loss vs ordinary SSIM loss, the MAE is enhanced by 3.5% and 5.4% on part A and part B. Above enhancement proves that Avg-MS-SSIM loss could help the network to learn the coarse-to-fined crowd features.

4 CONCLUSION

In this paper, we propose a Multi-Scale-Attention network called MSANet for crowd counting. From the perspective to make model could distinguish between foreground and background, a dual-stream network with attention map extractor and density map extractor is constructed. To enhance the model's ability to be aware of various scale information especially congested region features, we exploit dilated convolution in the back-end part of MSANet to enlarge the receptive field and a method similar to FPN to concatenate the same-sized feature map in the front and back-end network. Furthermore, to keep the local consistency of density map, a modified SSIM loss called Average Multi-Scale Structural Similarity loss (Avg MS-SSIM loss) is added to the final loss function. To handle the pedestrians counting problem in subway, we create a new subway dataset called SCCD with rich subway scenes. Experiments on ShanghaiTech Dataset and our SCCD demonstrate the powerful loss function settings and structure design of MSANet.

Table 6: Ablation Study of Pooling Times

Pooling times	MAE	MSE
2Avg MS-SSIM	60.46	95.05
3Avg MS-SSIM	**58.55**	**98.45**
4Avg MS-SSIM	58.66	98.34
5Avg MS-SSIM	62.61	104.59

Table 7: Ablation Study of Avg-MS-SSIM Loss on Shanghai Tech

Method	Part A		Part B	
	MAE	MSE	MAE	MSE
baseline	66.6	104.5	10.0	15.7
w. SSIM loss	60.6	102.5	7.4	12.3
w. Avg MS-SSIM loss	58.5	**98.45**	**7.0**	**11.1**

ACKNOWLEDGMENTS

This work is supported by National Key R&D Program of China under Grant 2020YFB160 0700, Major scientific research projects of China Railway Group under Grant K2019G046, the National Natural Science Foundation of China under grant 62001110, and the Natural Science Foundation of Jiangsu Province under grant SBK2020041044.

REFERENCES

[1] Viola P, Jones M J (2004). Robust real-time face detection. International journal of computer vision, pp. 137-154.
[2] Dalal, Navneet, and Bill Triggs (2005). Histograms of oriented gradients for human detection. IEEE computer society conference on computer vision and pattern recognition, pp. 886-893.
[3] Hu P, Ramanan D. (2017). Finding tiny faces. In Proceedings of the IEEE conference on. computer vision and pattern recognition, pp 951-959.
[4] Najibi M, Samangouei P, Chellappa R, *et al.* (2017). Ssh: Single stage headless face detector. Proceedings of the IEEE international conference on computer vision, pp. 4875-4884.
[5] Lempitsky V, Zisserman A. (2010). Learning to count objects in images. Advances in neural. information processing systems, pp 1324-1332.
[6] Zhang Y, Zhou D, Chen S, *et al.* (2016). Single-image crowd counting via multi-column convolutional. neural network. Proceedings of the IEEE conference on computer vision andpattern recognition, pp. 589-597.
[7] Li Y, Zhang X, Chen D. (2018). Csrnet: Dilated convolutional neural networks for understanding the. highly congested scenes. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1091-1100.
[8] Liu W, Salzmann M, Fua P. (2019). Context-aware crowd counting. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5099-5108.
[9] Lin T Y, Dollár P, Girshick R, *et al.* (2017). Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117-2125.
[10] Idrees H, Tayyab M, Athrey K, *et al.* (2018). Composition loss for counting, density map estimation and localization in dense crowds. Proceedings of the European Conference on Computer Vision, pp. 532-546.
[11] Rong L, Li C. (2021). Coarse-and fine-grained attention network with background-aware loss for crowd density map estimation. Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pp. 3675-3684.
[12] Shi M, Yang Z, Xu C, *et al.* (2019). Revisiting perspective information for efficient crowd counting. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7279-7288.
[13] Sindagi V A, Patel V M. (2019). Multi-level bottom-top and top-bottom feature fusion for crowd counting. Proceedings of the IEEE International Conference on Computer Vision, pp 1002-1012.
[14] Cheng Z Q, Li J X, Dai Q, *et al.* (2019). Improving the learning of multi-column convolutional neural network for crowd counting. Proceedings of the 27th ACM international conference on multimedia, pp. 1897-1906.
[15] Wei B, Yuan Y, Wang Q. (2020). MSPNET: multi-supervised parallel network for crowd counting. ICASSP-IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2418-2422.
[16] Wu X, Zheng Y, Ye H, *et al.* (2020). Counting crowds with varying densities via adaptive scenario discovery framework. Neurocomputing, pp. 127-138.
[17] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. (2018). Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141.
[18] Zhao Z, Han T, Gao J, *et al.* (2020). A flow base bi-path network for cross-scene video crowd understanding in aerial view. European Conference on Computer Vision. Springer, Cham, pp. 574-587.
[19] Valloli V K, Mehta K. (2019). W-net: Reinforced u-net for density map estimation[J]. arXiv preprint arXiv. pp. 1903.11249.
[20] Ronneberger O, Fischer P, Brox T. (2015). U-net: Convolutional networks for biomedical image. Segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, pp. 234-241.
[21] Shi Z, Mettes P, Snoek C G M. (2019).Counting with focus for free. Proceedings of the IEEE International Conference on Computer Vision, pp. 4200-4209.
[22] Lin T Y, Goyal P, Girshick R, *et al.* (2017). Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision, pp. 2980-2988.
[23] Liu N, Long Y, Zou C, *et al.* (2019). ADcrowdNet: An attention-injective deformable convolutional network for crowd understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3225-3234.
[24] Cao X, Wang Z, Zhao Y, *et al.* (2018) Scale aggregation network for accurate and efficient crowd counting. Proceedings of the European Conference on Computer Vision, pp. 734-750.
[25] Liu L, Qiu Z, Li G, *et al.* (2019). Crowd counting with deep structured scale integration network. Proceedings of the IEEE International Conference on Computer Vision, pp. 1774-1783.
[26] Wang Q, Gao J, Lin W, *et al.* (2019). Learning from synthetic data for crowd counting in the wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8198-8207.
[27] He G, Chen Q, Jiang D, *et al.* (2017). A double-region learning algorithm for counting the number of pedestrians in subway surveillance videos[J]. Engineering Applications of Artificial Intelligence, pp. 302-314.
[28] Jiang S, Lu X, Lei Y, *et al.* (2019). Mask-aware networks for crowd counting[J]. IEEE Transactions on. Circuits and Systems for Video Technology, 30(9): 3119-3129.
[29] Zhu L, Zhao Z, Lu C, *et al.* (2019). Dual path multi-scale fusion networks with attention for crowd. counting[J]. arXiv preprint arXiv:1902.01115.
[30] Ioffe S, Szegedy C. (2015). Batch normalization: Accelerating deep network training by reducing. internal covariate shift. International conference on machine learning. PMLR, pp448-456.
[31] Jiang X, Xiao Z, Zhang B, *et al.* (2019). Crowd counting and density estimation by trellis encoder-decoder networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6133-6142.