symmetry

Article

Convolutional Neural Network for Crowd Counting on Metro Platforms

Jun Zhang [1], Jiaze Liu [1] and Zhizhong Wang [2,*]

[1] School of Mechanical and Power Engineering, Zhengzhou University, Zhengzhou 450000, China; zhangjun@zzu.edu.cn (J.Z.); ljz_0075@gs.zzu.edu.cn (J.L.)
[2] School of Electrical Engineering, Zhengzhou University, Zhengzhou 450000, China
* Correspondence: wzz1982@zzu.edu.cn

Abstract: Owing to the increased use of urban rail transit, the flow of passengers on metro platforms tends to increase sharply during peak periods. Monitoring passenger flow in such areas is important for security-related reasons. In this paper, in order to solve the problem of metro platform passenger flow detection, we propose a CNN (convolutional neural network)-based network called the MP (metro platform)-CNN to accurately count people on metro platforms. The proposed method is composed of three major components: a group of convolutional neural networks is used on the front end to extract image features, a multiscale feature extraction module is used to enhance multiscale features, and transposed convolution is used for upsampling to generate a high-quality density map. Currently, existing crowd-counting datasets do not adequately cover all of the challenging situations considered in this study. Therefore, we collected images from surveillance videos of a metro platform to form a dataset containing 627 images, with 9243 annotated heads. The results of the extensive experiments showed that our method performed well on the self-built dataset and the estimation error was minimum. Moreover, the proposed method could compete with other methods on four standard crowd-counting datasets.

Keywords: metro platform; crowd counting; multiscale feature extraction; convolutional neural network

check for updates

Citation: Zhang, J.; Liu, J.; Wang, Z. Convolutional Neural Network for Crowd Counting on Metro Platforms. *Symmetry* **2021**, *13*, 703. https://doi.org/10.3390/sym13040703

Academic Editor: Raúl Baños Navarro

Received: 19 March 2021
Accepted: 13 April 2021
Published: 17 April 2021

1. Introduction

Owing to the rapid development of urban rail transit, the lines of operation are expanding, passenger flow continues to increase [1], and rail operators face daunting safety-related challenges in this context. Crowd density in metro stations increases sharply in peak periods of travel. As a large crowd gathers at metro stations and passenger flows increase, the risk of stampedes increases. Therefore, it is important to analyze passenger flow by monitoring videos of the metro platform, analyzing their content, and identifying abnormalities using computer vision and artificial intelligence [2,3]. According to information on real-time passenger flows and crowd densities in different areas, people on a platform can be guided to avoid stampedes, improving the security and efficiency of metro stations.

A considerable amount of research has been conducted on analyzing the flow of passengers through metro stations based on surveillance videos [2,4,5]. In [2], passenger flow in a given target area was detected using the background difference method, but this method cannot be used to count the number of passengers on the metro platform. Background difference is more suitable for the detection of continuously moving objectives; passengers waiting on the metro platform are mostly stationary. The authors of [4] proposed a strategy to detect passenger flow on a metro platform based on the bodies of the passengers. In a sparse scene, this method performs well, but the metro platform is highly crowded at times, as shown in Figure 1. In such cases, images captured by the camera feature significant occlusions that cause this method to miss some targets and incorrectly identify others. In [5], the authors proposed a crowd monitoring approach for metro platforms using an improved mixture of Gaussian background modeling to segment the crowd. People in the crowd are counted by linear regression. This method regards the crowd as a whole and uses the regression relationship between features of the image and the crowd to count the passengers. It can solve the problem of occlusion and count people in a large crowd. However, the accuracy of this method is low owing to the limited population information provided by the crowd.

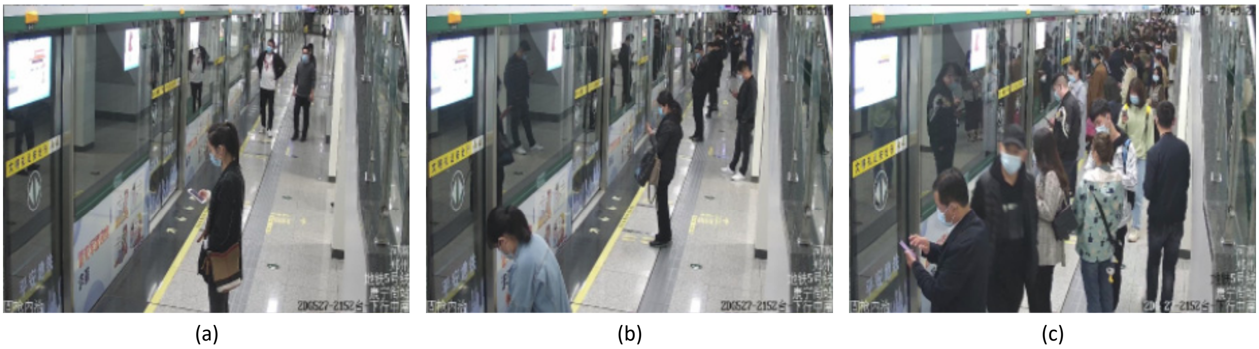

Figure 1. Representative images in the Metro Platform dataset, (**a**–**c**) represent different crowding degrees.

Crowd counting methods aim to estimate the number of people in surveillance videos or a single image. They can be used in a variety of scenarios, such as political assemblies, sports events, and concerts, to ensure public safety by monitoring crowd density. Currently available methods of crowd counting are developed from detection-based [4,6–8] and regression-based [5,9,10] approaches and convolutional neural network (CNN)-based [11–16] approaches. As CNN-based methods use the human head as the target of detection, the error caused by occlusion is reduced. Therefore, convolutional-neural-network-based crowd counting methods are more suitable for use on metro platforms.

The presence of screen doors, elevators, and other small facilities on metro platforms, as well as changes in lighting, can cause severe occlusion and reflection problems in surveillance videos captured by monitoring probes. This seriously affects the accuracy of crowd counting. The problems of occlusion and reflection pose significant challenges to crowd counting at metro platforms. To solve the occlusion and reflection problem, we propose a convolutional neural network for crowd counting called the MP-CNN. The proposed architecture uses VGG-16 [17] as the front-end network for feature extraction. The VGG is known to have excellent feature extraction capability and strong transfer learning ability on classification tasks. It also has flexible architecture, which makes it easy to connect it to the back-end network and generate a density map. Inspired by the work in [18], we also introduce a multiscale feature extraction module to enhance the multiscale feature extraction capability and expand the field of reception of the network. This can improve feature extraction in remote areas of a long and narrow metro platform. We then use a set of transposed convolutions for upsampling, instead of bilinear interpolation, to restore the feature map to its original size and generate a high-quality density map.

We also developed a dataset that contains 627 images of a total of 9243 annotated people for this study. It contains images of a platform at peak and normal periods on weekdays and weekends. Owing to the long and narrow metro platform considered and the angles of the surveillance cameras, the degrees of crowding and occlusion were different. The data were collected from a surveillance video camera on the metro platform. We call it the Metro Platform dataset. The representative images of the proposed dataset are shown in Figure 1.

The main contributions of our work are as follows:

First, for the sake of public safety, in order to avoid stampede accidents, we propose a convolutional neural network called the MP–CNN for accurate crowd counting on metro platforms.

Second, the proposed method, with a multiscale feature extraction module, can solve the problem of severe occlusion and better adapt to environments with severe occlusion and reflection compared to other methods.

Third, we developed a Metro Platform dataset; images in this dataset were gathered from a video stream of a metro station. This dataset has different scenes featuring congestion for analysis in the field of intelligent transportation.

The results of experiments on the four benchmarks show that our method can compete with state-of-the-art crowd counting methods.

2. Related Work

In recent years, a growing number of studies have considered the problem of crowd counting and proposed algorithms to deal with this task. They can be broadly categorized into traditional methods and CNN-based methods.

2.1. Traditional Approach

Early work on crowd counting focused on detection-based methods [6,19–22]. Some of them considered the crowd as a group of detected individual pedestrians by using a simple process of detection and summation. Others tackled crowd counting as an object detection problem and used the body, or parts of it, to locate people in images of crowds in order to count them. However, in scenes of dense crowds, these detection-based methods were limited by serious occlusion and background clutter. To handle images of highly congested scenes, regression-based approaches [9,10,23,24] were proposed. They involved learning to map from features of the image to density maps or to a given number of particular objects directly. Using similar approaches, in [24], Idrees et al. proposed a method that fuses the extracted features using Fourier analysis, head detection, and scale-invariant feature transform in local patches. These regression-based methods can predict the global number of people in a crowd but ignore the spatial information in images. A comprehensive survey of these early studies can be found in [25].

2.2. CNN-Based Approach

Various CNN-based methods have been proposed and have achieved remarkable success in crowd counting tasks. A majority of them are dedicated to large-scale variations in images of crowds. The authors of [26,27] have summarized the previously proposed CNN-based methods for crowd counting. To cope with the large-scale variation in scenes of crowds, Zhang et al. [12] proposed a simple and effective multicolumn structure to extract features by kernel size. Similarly, in [28], a multiscale model, hydra-CNN, was proposed by Onoro and Sastre to extract image features at different scales. Cao et al. [13] proposed an encoder–decoder network called SANet that employs scaled aggregation modules as an encoder. This method can improve representation capability and the diversity of feature scale. Recently, Wang [15] designed a network called SFCN to encode spatial contextual information based on the VGG-16 [17] and ResNet-101 [29]. The problem of scale variation can be solved by certain techniques, such as dilated kernels [14], multiscale pooling layers [30], multiple decoding paths [31], and multiscale bottom-up and top-down feature fusion [32].

The above studies show that CNN-based solutions can outperform traditional methods of crowd counting. We thus propose a CNN-based network, with pooling layers and dilated convolution [14], to solve the problem, as applies to a metro platform.

3. Proposed Method

In this section, we first introduce the architecture of the proposed convolutional neural network for crowd counting on metro platforms (MP-CNN), as shown in Figure 2. We then discuss the multiscale feature extraction module (MFEM) and the method for generating ground truth. Finally, we describe details of the training of the proposed method.

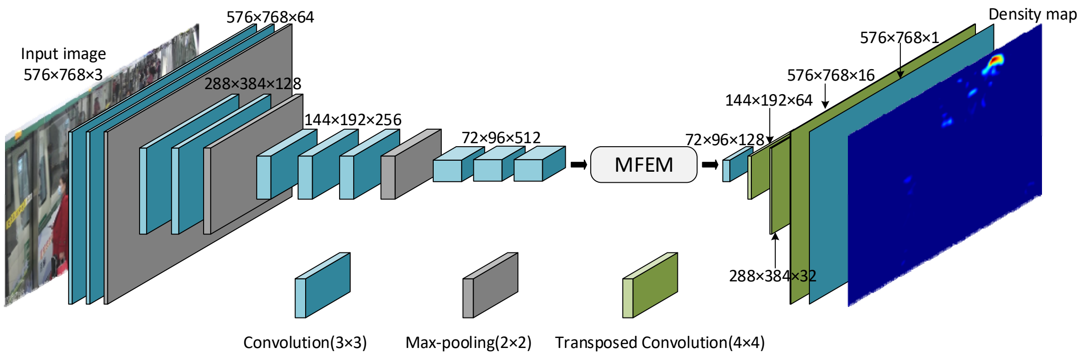

Figure 2. The architecture of the proposed metro platform convolutional neural network (MP-CNN).

3.1. Architecture

We use the first 13 layers of the VGG-16 [17] as the front-end network for feature extraction and only a 3 × 3 convolution kernel. We chose the VGG as the front end for two reasons. On the one hand, it has excellent feature extraction capability and a strong transfer learning ability for classification tasks; on the other hand, the VGG has flexible architecture, which makes it easy to connect to the back-end network to generate a density map. After a series of convolution layers and pooling layers in the front-end network, the size of the output feature map is 1/8 of the original input. If we continue to stack more convolution layers and pooling layers, the size of the output feature map can be further reduced, and it becomes difficult to generate a high-quality density map. Therefore, after processing at the front end, we introduced the MFEM, which can extract deeper information while maintaining the resolution of the output density map. The dilated convolution shown in Figure 3b is used in this module. Dilated convolutional layers are known to significantly improve predictive accuracy on semantic segmentation tasks [33,34].

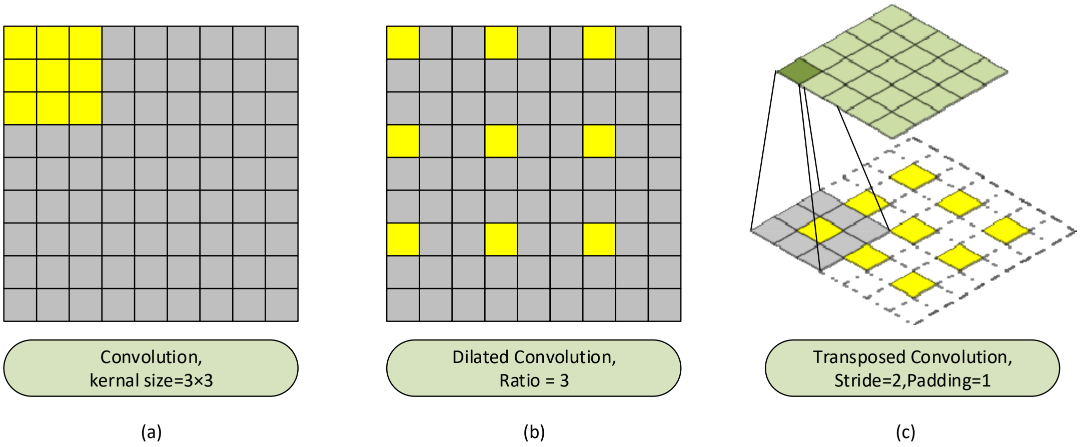

Figure 3. Comparison of three convolutions.

Because of the downsampling of the image in the feature extraction process, the resolution of the output feature is reduced, and it loses considerable detail. To obtain a high-resolution density map, we use a set of transposed convolutions to upsample the image after the MFEM has been used. A transposed convolution is not a completely inverse process of a normal convolution but a special convolution. The image size is first expanded by padding the image with 0 s according to a certain ratio. The convolution kernel is then rotated, and forward convolution is performed, as shown in Figure 3c. Unlike previous methods, we chose a learnable transposed convolution instead of a bilinear interpolation algorithm for upsampling. Transposed convolution is different from bilinear interpolation in that it has parameters that can be learned, which means that it can learn more feature information than bilinear interpolation. The transposed convolution layers are used to restore the spatial resolution of the image. Each transposed convolution layer doubles the size of the feature map, corresponding to the previous max-pooling layer. Three transposed convolution layers are used in the network to generate a high-resolution density map of the same size as the input image. This provides detailed spatial information to facilitate feature learning while training the model.

3.2. Multiscale Feature Extraction Module

Owing to the complex distribution of passengers waiting on a metro platform, the perspective of the camera, and other problems, the head size of passengers in the captured images varies. In addition, reflections from screen doors on the platform, elevators, and other small facilities cause complex changes in background information. These problems pose daunting challenges to the crowd counting task on the metro platform. Previously proposed methods, such as the L2SM [35] and S-DCNet [36], have focused on fusing feature maps from different CNN layers to acquire multiscale information through a feature pyramid network structure. In this paper, we introduce a multiscale feature extraction module to solve this problem. This is the first time we have applied this method to the crowd counting task of a metro platform. The proposed MFEM improves multiscale feature extraction to enhance the information in each layer of the feature map.

As shown in Figure 4, the MFEM first compresses the channel of the feature map via a 1×1 convolution and then processes the compressed feature map by dilated convolution, with different dilated ratios of 1, 2, 3, and 4 to handle the multiscale features and variations in head sizes in the images. The size of the fixed Gaussian kernel in this paper is set to 15. In the generated density map, the size of each annotated head is 15×15; padding the image with some 0 s does not affect the counting result. Dilated convolution expands the receptive field of the convolution kernel while keeping the number of parameters unchanged; the operation speed can be accelerated by doing this. The diagrammatic sketch of dilated convolution is shown in Figure 3b, of which the dilated ratio is 3. The extracted multiscale feature maps are fused by the concatenation operation and a 3×3 convolution; the size of the processed feature images is the same as that of the input images.

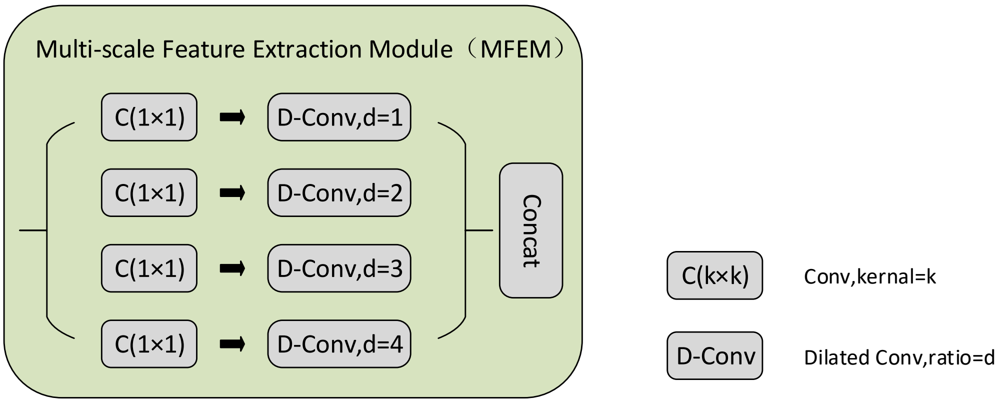

Figure 4. Structure of the multiscale feature extraction module.

The key component of this design is the dilated convolution layer. A dilated convolution can be defined as follows:

$$Y(l,w) = \sum_{i=1}^{L}\sum_{j=1}^{W} x(l + d \times i, w + d \times j) f(i,j) \tag{1}$$

$Y(l,w)$ is the output of the dilated convolution from input $x(l,w)$. Filter $f(i,j)$ has the length and width L and W, respectively. Parameter d represents the rate of dilation. When $d = 1$, a dilated convolution turns into a normal convolution.

3.3. Ground Truth Generation

In research on crowd counting, the dataset used is typically composed of original images and annotated files. Annotations for images of crowds include points at the center of each passenger's head, which record the two-dimensional (2D) coordinates of each head and the total number of heads. This is required to convert these discrete coordinate points into a density map to predict passenger density.

The ground-truth density map is generated by convolving each delta function $\delta(x - x_i)$ with a normalized Gaussian kernel G_σ:

$$F = \sum_{i=1}^{N} \delta(x - x_i) * G_\sigma \tag{2}$$

where x represents each pixel in a given image, x_i is the ith annotated point, and N is the set of all annotated points. The integral of the density map is equal to the number of people in the image. Instead of using geometry-adaptive kernels, as in [12], we use a fixed Gaussian kernel to generate the ground-truth density maps; the spread parameter σ of the Gaussian kernel is set to 15.

The sum of all pixel values gives the number of people in the crowd in the input image. P denotes the number of passengers and is defined as follows:

$$P = \sum_{l=1}^{L} \sum_{w=1}^{W} Z_{l,w} \tag{3}$$

where L represent the length of the density map and W represents the width of the density map. Moreover, $Z_{l,w}$ is the pixel at (l, w) in the generated density map.

3.4. Training Details

We trained the proposed MP-CNN in an end-to-end manner. The weight parameters of the VGG net, trained on ImageNet, were used for pretraining. We perform our experiments on an NVIDIA Quadro P4000 GPU, with batch size = 1. An Adam optimizer [37] with a low learning rate of 1×10^{-5} was used to train the model; all experiments are trained for 500 epochs. The Euclidean distance was used to measure estimation error at the pixel level, as in [12,14,28]. The loss function was defined as follows:

$$L(\theta) = \frac{1}{2N} \sum_{i=1}^{N} \| F(X_i; \theta) - F_i \|_2^2 \tag{4}$$

In the above equation, θ denotes a set of parameters in the proposed MP-CNN, N is the number of training images, X_i represents the input image, and F_i denotes the ground-truth density map of image X_i. $F(X_i; \theta)$ stands for the estimated density map generated by the MP-CNN, parameterized with θ for the sample, and L is the loss between the estimated density map and the ground-truth density map. Our method was implemented on the Pytorch [38] framework.

4. Experiments

In this section, we first introduce the datasets and evaluation metrics used. The experiments conducted on the Metro Platform dataset are then detailed. They verify that the proposed method can be used for counting passengers on a metro platform. We then compare our method with state-of-the-art methods on four standard datasets to prove its generalization capability. Finally, we report ablation studies to prove the effectiveness of the proposed MFEM used in our method.

4.1. Datasets

We evaluated our method on four publicly available crowd counting benchmark datasets as well as the dataset collected for this paper (Metro Platform): ShanghaiTech [12] Part A and Part B, UCF-QNRF [39], and UCF-CC-50 [24].

ShanghaiTech. The ShanghaiTech dataset was developed by [12] and contains 1198 images, with 330,165 annotated people. Each image in this dataset has a different perspective. This dataset consists of two parts: Part A with 482 images and Part B with 716 images. The crowd density varies significantly between Part A and Part B, making the accurate estimation of the crowd more challenging. Images in Part A were randomly collected from the internet, and Part B contains images captured from street views. We used the training and testing set splits provided by the authors; in this way, we had 300 images for training and 182 images for testing in Part A and 400 images for training and 316 images for testing in Part B.

UCF-QNRF. As we all know, UCF-QNRF is the largest and most widely distributed dataset in the domain of crowd counting, reported in [39] in 2018. This dataset contains 1535 images featuring 1,251,642 people, with the centers of their heads annotated, including 1201 images in the training set and 334 images in the test set. A wide variety of scenes are contained, including a diverse set of viewpoints, densities, and variations in lighting. The resolution is higher than in the ShanghaiTech dataset. This makes this dataset more realistic as well as more difficult when counting the number of people in the image.

UCF-CC-50. The UCF-CC-50 dataset [24] contains 50 annotated images of extremely dense crowds. The images were collected mainly from concerts, protests, and marathons, with different crowd densities and perspectives. There is a large variation in crowd numbers, ranging from 94 to 4543. The limited number of images makes it a challenging dataset for deep learning methods. We followed the standard protocol in [24] and used five-fold cross-validation to evaluate the performance of the proposed method on this dataset.

Metro Platform. Crowd counting is important, but the available counting datasets are not specifically designed for metro transportation. Therefore, we collected and labeled a dataset that is specific to metro platforms in order to count the waiting passengers in such areas. The images were captured from a video from a camera at a certain perspective on a metro platform, including the peak and normal periods on weekdays and weekends. The Metro Platform dataset consists of 627 images and 9243 annotations; the resolution of the images is 576 × 768. For the evaluation, we used 465 images from the dataset as the training set and 162 as the testing set. A comparison between the Metro Platform dataset and the other datasets used is shown in Table 1.

Table 1. Comparison between the Metro Platform dataset and the other datasets used in this study: Num is the number of images, Total is the total number of labeled people, Ave is the average crowd count, and Max is the maximal crowd count.

Dataset	Num	Average Resolution	Annotations		
			Total	Ave	Max
ShanghaiTech_PartA [12]	482	589 × 868	241,677	501	3139
ShanghaiTech_PartB [12]	716	768 × 1024	88,488	123	578
UCF-QNRF [39]	1535	2013 × 2902	1,251,642	815	12,865
UCF-CC-50 [24]	50	2101 × 2888	63,974	1279	4633
Metro Platform	627	576 × 768	9243	15	43

4.2. Evaluation Metrics

In accordance with previous studies [12–14], we used mean absolute error (MAE) and mean squared error (MSE) as metrics to evaluate the accuracy of the methods in terms of counting members of a crowd:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Z_i - \overset{\wedge}{Z_i}\right| \tag{5}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Z_i - \overset{\wedge}{Z_i})^2} \tag{6}$$

In the above equation, N is the number of test samples; Z_i and $\overset{\wedge}{Z_i}$ are the estimated and ground-truth crowd numbers corresponding to the ith sample, which is given by the integration of the density map. Roughly speaking, the MAE indicates the accuracy of the predicted result and the MSE measures its robustness. As the MSE is sensitive to outliers, its value will be large when the model performs poorly on a few samples.

4.3. Experiments on the Metro Platform Dataset

The Metro Platform dataset was designed specifically for metro platforms. Due to the angle of the camera, the characteristics of the crowd close to the camera are clear, while those at a long distance from it are blurred, as shown in Figure 1. In addition, the background in the image is more complex, and background information accounts for a large part of the image. The screen door of the metro platform also produces significant reflection. The position of the crowd in each image changes, and the adverse background caused by the reflection also changes. The above problems pose significant challenges for the counting task. To solve the problem of changeable background, we introduced the model trained on the dense crowd datasets as a pretrained model in the experiments. We used the model trained on ShanghaiTech Part A as the pretrained model to evaluate network performance. The results of the comparison are shown in Table 2. Figure 5 shows the density map obtained using the different methods. We adopted memory access cost (MAC) to evaluate the computational complexity. In the same experimental environment, the MAC values of different methods are shown in Table 2. Our method achieves the highest counting accuracy on metro platform scenes, but the network is also more complex. In future work, we will try to lightweight the network.

Table 2. Comparing performances of different methods on the Metro Platform dataset. MAC is the memory access cost.

Method	MAE	MSE	MAC (GB)
MCNN [12]	2.6	3.3	1.351
CSRNet [14]	3.2	4.1	20.737
SANet [13]	2.9	4.1	4.559
Ours (with MFEM)	1.6	2.2	26.771

The proposed method was superior to the other methods in the metro platform scenario as it was more accurate in terms of counting the number of passengers in crowds and generated a higher-quality density map. The distribution of passengers on the metro platform can be obtained from the generated density map, and subway staff can dredge the crowd in the crowded area according to the actual situation so as to avoid safety accidents caused by overcrowding in a certain area.

In the future, we will further explore the influence of occlusion and reflection on the counting task of metro platforms. We will try to improve the estimation accuracy in two different ways. First, use hybrid supervised–unsupervised machine learning approaches [40,41] in an attempt to extract more relevant features. Second, preprocess the monitoring video to cut or cover the screen door of the metro platform that had a serious problem of reflection.

4.4. Comparisons with State of the Art

The proposed method delivered outstanding performance on all benchmarks. The results of quantitative comparisons with the state-of-the-art methods on four datasets are presented in Tables 3 and 4. A visual comparison is also provided in Figure 6.

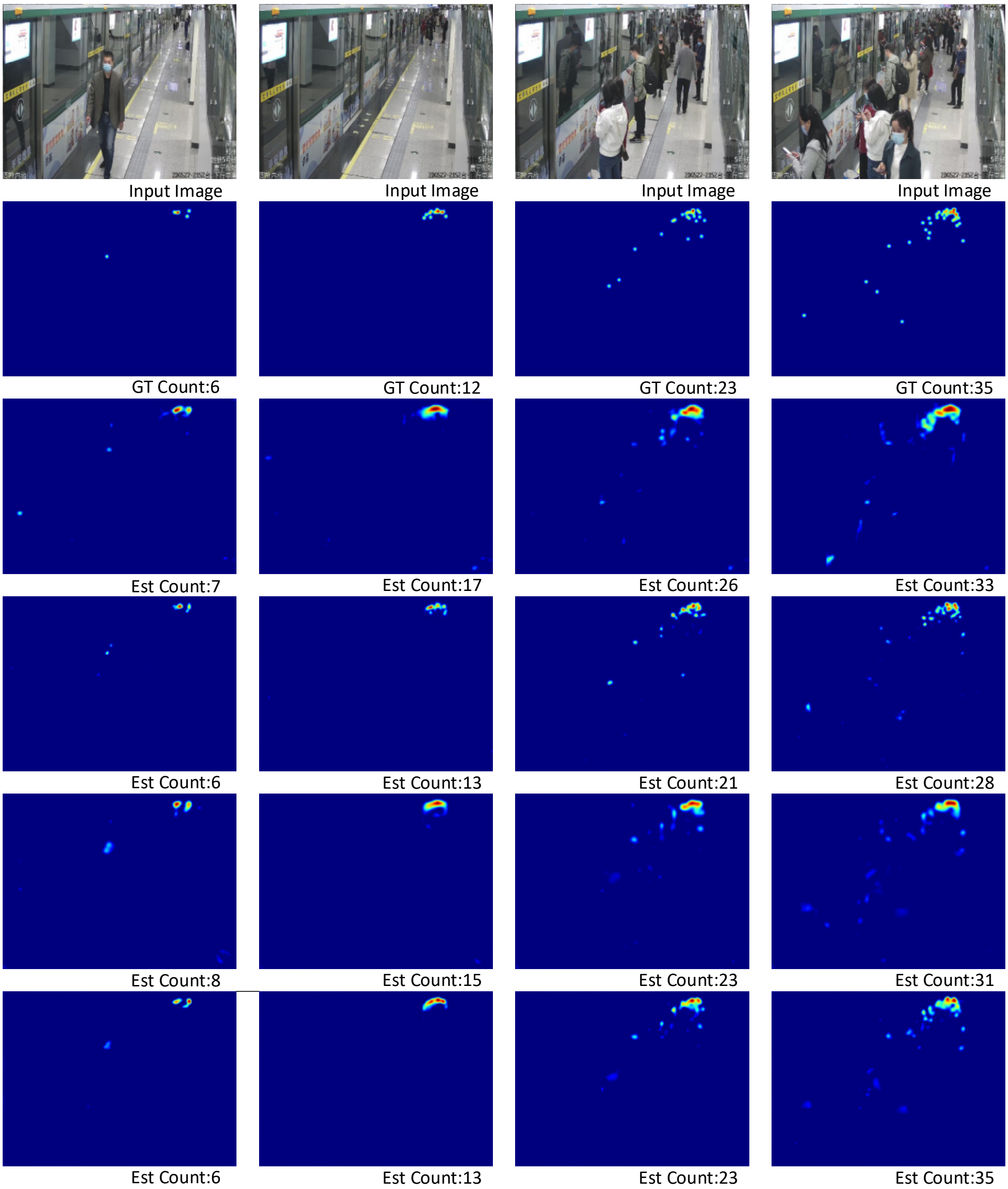

Figure 5. Visualization of estimated density maps on the Metro Platform dataset. First row: input images. Second row: ground truth. Third row: density maps estimated by MCNN. Fourth row: density maps estimated by SANet. Fifth row: density maps estimated by the proposed method without MFEM. Sixth row: density maps estimated by the proposed method with MFEM.

ShanghaiTech. We compared the proposed method with multiple classic methods on ShanghaiTech Part A and Part B datasets and found that it yielded a significant improvement in performance. In Part A, our method was superior by 39.2% in terms of the MAE, 34.9% in terms of the MSE to the MCNN, and, respectively, by 1.8% and 2.1% to CSRNet. In Part B, our method was superior by 62.5% in terms of the MAE, 64.6% in terms of the MSE compared to the MCNN, and by 6.6% and 8.8% to CSRNet, respectively.

UCF-QNRF. As we all know, UCF-QNRF is the largest and most widely distributed crowd counting dataset. The proposed method achieved significant improvement over existing methods on this dataset. For instance, RANet [42] achieved a score of 111 in terms of the MAE and 190 in terms of the MSE, whereas our method improved these results by 3.3% in terms of the MAE and 4.3% in terms of the MSE.

UCF-CC-50. We also conducted experiments on the UCF-CC-50 dataset. The crowd numbers in the images varied from 96 to 4633. According to the standard protocol in [24], the dataset was randomly divided into five subsets. We used five-fold cross-validation to evaluate our method. With a small number of training images, our network still converged well on this dataset. Compared with RANet [42], it was better by 5.5% in terms of the MAE and 4.3% in terms of the MSE.

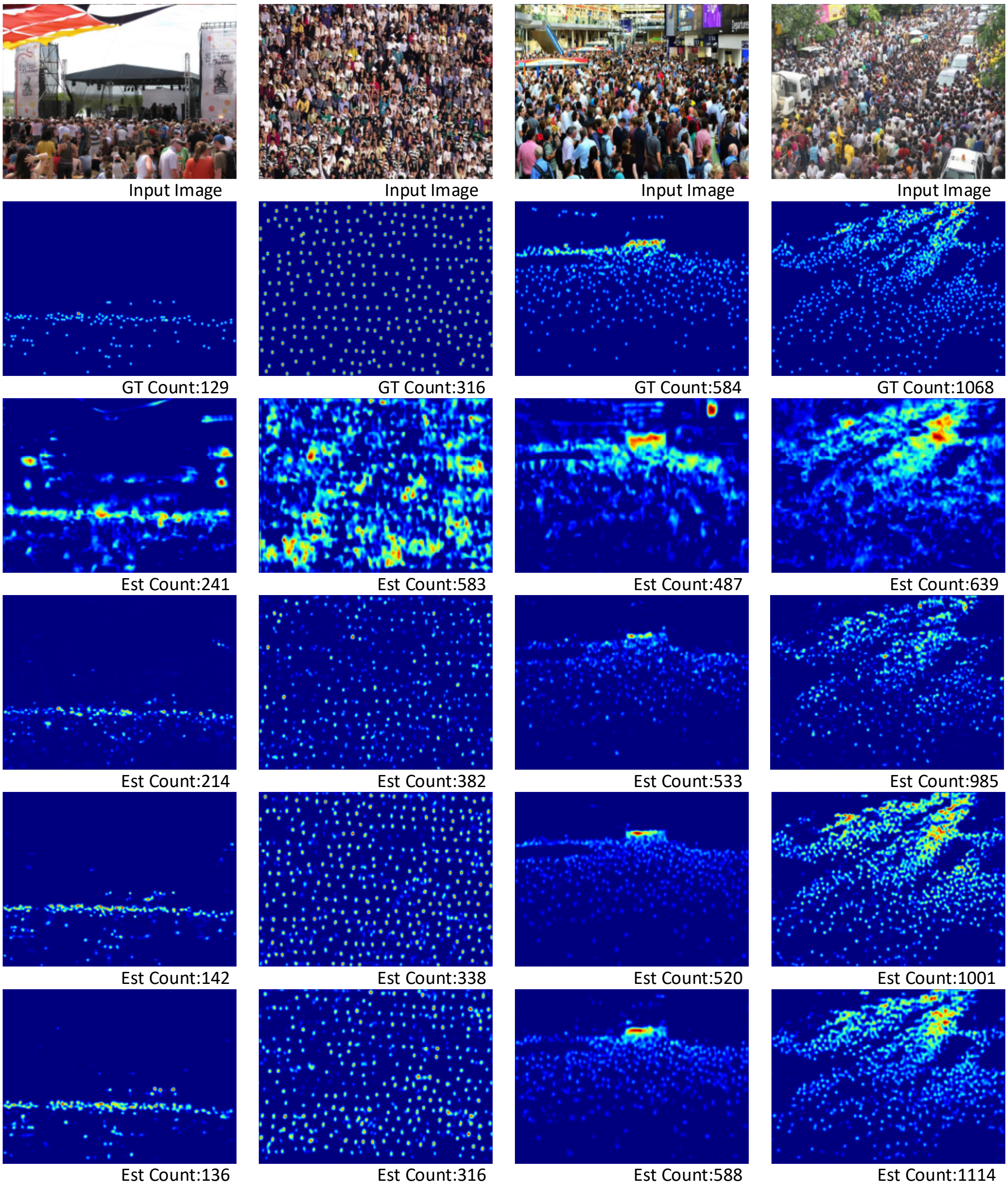

Figure 6. Visualization of estimated density maps on the ShanghaiTech Part A dataset. First row: input images. Second row: ground truth. Third row: estimated density generated maps by MCNN. Fourth row: estimated density maps generated by CSRNet. Fifth row: estimated density maps generated by the proposed method without MFEM. Sixth row: estimated density maps generated by the proposed method with MFEM.

Table 3. Comparing results of different methods on the ShanghaiTech Part A and Part B datasets.

Method	Part A		Part B	
	MAE	MSE	MAE	MSE
MCNN [12]	110.2	173.2	26.4	41.3
Switch-CNN [43]	90.4	135.0	21.6	33.4
L2R [44]	73.6	112.0	13.7	21.4
IG-CNN [45]	72.5	118.2	13.6	21.1
CSRNet [14]	68.2	115.0	10.6	16.0
CP-CNN [46]	73.6	106.4	20.1	30.1
ic-CNN [47]	68.5	116.2	10.7	16.0
Ours (with MFEM)	67.0	112.6	9.9	14.6

Table 4. Comparing results of different methods on UCF-QNRF and UCF-CC-50 datasets.

Method	UCF-QNRF		UCF-CC-50	
	MAE	MSE	MAE	MSE
MCNN [12]	277	426	377.6	509.1
Switch-CNN [43]	228	445	318.1	439.2
Encoder-Decoder [31]	113	188	249.4	354.5
RANet [42]	111	190	239.8	319.4
CSRNet [14]	-	-	266.1	397.5
Composition Loss [39]	132	191	-	-
Ours (with MFEM)	107.3	181.8	226.5	305.6

4.5. Ablation Experiments

In this section, we report the results of ablation studies on the different datasets used to verify the effectiveness of the proposed MFEM. The experimental results show that while considering the balance of training speed and estimation accuracy, the structure setting of MFEM in Figure 4 is the best choice.

To verify the effectiveness of the MFEM, we used the proposed network structure with MFEM and without it in the training process on different datasets. The results showed that the performance of the proposed method improved when the MFEM was introduced, as shown in Table 5. On the ShanghaiTech Part B dataset, the proposed MFEM improved the performance by 21.4% in terms of the MAE and 25.9% in terms of the MSE. On the UCF-CC-50 dataset, the introduction of MFEM improved the performance by 12.7% and 5.6% in terms of the MAE and MSE, respectively. On the Metro Platform dataset proposed in this paper, the MAE and MSE improved by 33.3% and 29%, respectively, with the introduction of MFEM. This shows that the MFEM can improve counting performance in dense and relatively sparse scenes.

Table 5. Ablation study on the multiscale feature extraction module.

Dataset	Without MFEM		With MFEM	
	MAE	MSE	MAE	MSE
ShanghaiTech Part A [12]	67.2	119.0	67.0	112.6
ShanghaiTech Part B [12]	12.6	19.7	9.9	14.6
UCF-QNRF [39]	110.5	182.1	107.3	181.8
UCF-CC-50 [24]	259.5	323.7	226.5	305.6
Metro Platform	2.4	3.1	1.6	2.2

5. Conclusions

In this paper, we propose a novel method to count the number of people in crowds on metro platforms, called the MP-CNN. We introduced an MFEM to enhance the multiscale feature extraction capability of the network and solved the problems of diverse occlusion and varying head sizes of passengers in the images. This method is of great significance to the public safety of metro platforms; metro staff can guide and drain the flow according to the number of passengers. The effectiveness of the proposed MFEM was verified by comparative experiments. To evaluate its effectiveness on metro platforms in particular, we collected and labeled a new dataset, called the Metro Platform dataset, consisting of 627 images of 9243 annotated people. The results of extensive experiments show that our method delivers excellent results on the proposed Metro Platform dataset and can compete with state-of-the-art methods in four major crowd counting benchmarks.

Author Contributions: Resources, conceptualization, J.Z. and Z.W.; software, methodology, formal analysis, validation, data curation, investigation, writing—original draft, writing—review and editing, visualization, J.L.; funding acquisition, supervision, project administration, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC) General Program, grant number 61673353.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank Zhengzhou Metro Group Co. Ltd. for the original data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. China Urban Rail Transit Association. Urban Rail Transit 2018 Annual Statistics and Analysis Report. *Urban Rail Transit* **2019**, *4*, 16–34.
2. Qian, X.; Yu, X.; Fa, C. The passenger flow counting research of subway video based on image processing. In Proceedings of the 2017 29th Chinese Control And Decision Conference (CCDC), Chongqing, China, 28–30 May 2017; pp. 5195–5198.
3. Chato, P.; Chipantasi, D.J.M.; Velasco, N.; Rea, S.; Hallo, V.; Constante, P. Image processing and artificial neural network for counting people inside public transport. In Proceedings of the 2018 IEEE Third Ecuador Technical Chapters Meeting (ETCM), Cuenca, Ecuador, 15–19 October 2018; pp. 1–5.
4. Sheng, Z.; Tian, K.; Tian, Q.; Qu, H. A Faster R-CNN Based High-Normalization Sample Calibration Method for Dense Subway Passenger Flow Detection. In Proceedings of the 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Beijing, China, 13–15 October 2018; pp. 1–5.
5. Hu, X.; Zheng, H.; Wang, W.; Li, X.; Optics, E. A novel approach for crowd video monitoring of subway platforms. *Optik* **2013**, *124*, 5301–5306. [CrossRef]
6. Li, M.; Zhang, Z.; Huang, K.; Tan, T. Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection. In Proceedings of the 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008.
7. Ge, W.; Collins, R.T. Marked point processes for crowd counting. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 2913–2920.
8. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 743–761. [CrossRef] [PubMed]
9. Lempitsky, V.; Zisserman, A. Learning to count objects in images. In Proceedings of the Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, Vancouver, BC, Canada, 6–9 December 2010.
10. Pham, V.Q.; Kozakaya, T.; Yamaguchi, O.; Okada, R. COUNT Forest: CO-Voting Uncertain Number of Targets Using Random Forest for Crowd Density Estimation. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
11. Zhang, C.; Li, H.; Wang, X.; Yang, X. Cross-scene crowd counting via deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 833–841.
12. Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; Ma, Y. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
13. Cao, X.; Wang, Z.; Zhao, Y.; Su, F. Scale aggregation network for accurate and efficient crowd counting. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.

14. Li, Y.; Zhang, X.; Chen, D. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
15. Wang, Q.; Gao, J.; Lin, W.; Yuan, Y. Learning from Synthetic Data for Crowd Counting in the Wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
16. Zhang, J.; Zhu, G.; Wang, Z.J.S. Multi-Column Atrous Convolutional Neural Network for Counting Metro Passengers. *Symmetry* **2020**, *12*, 682. [CrossRef]
17. Simonyan, K.; Zisserman, A.J.C.S. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
18. Liu, X.; Yang, J.; Ding, W. Adaptive Mixture Regression Network with Local Counting Map for Crowd Counting. *arXiv* **2020**, arXiv:2005.05776.
19. Lin, S.F.; Chen, J.Y.; Chao, H.X. Estimation of number of people in crowded scenes using perspective transformation. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **2001**, *31*, 645–654.
20. Lin, Z.; Davis, L.S. Shape-Based Human Detection and Segmentation via Hierarchical Part-Template Matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 604–618. [CrossRef] [PubMed]
21. Wang, M.; Wang, X. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In Proceedings of the CVPR, Colorado Springs, CO, USA, 20–25 June 2011.
22. Wu, B.; Nevatia, R. Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors. *Int. J. Comput. Vision* **2007**, *75*, 247–266. [CrossRef]
23. Chen, K.; Loy, C.C.; Gong, S.; Xiang, T. Feature Mining for Localised Crowd Counting. *Bmvc* **2012**, *1*, 3.
24. Idrees, H.; Saleemi, I.; Seibert, C.; Shah, M. Multi-Source Multi-Scale Counting in Extremely Dense Crowd Images. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013.
25. Li, T.; Chang, H.; Wang, M.; Ni, B.; Hong, R.; Yan, S. Crowded Scene Analysis: A Survey. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *25*, 367–386. [CrossRef]
26. Sindagi, V.A.; Patel, V.M. A Survey of Recent Advances in CNN-based Single Image Crowd Counting and Density Estimation. *Pattern Recognit. Lett.* **2018**, *107*, 3–16. [CrossRef]
27. Gao, G.; Gao, J.; Liu, Q.; Wang, Q.; Wang, Y. CNN-based Density Estimation and Crowd Counting: A Survey. *arXiv* **2020**, arXiv:2003.12783.
28. Ooro-Rubio, D.; López-Sastre, R.J. Towards Perspective-Free Object Counting with Deep Learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2016.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
30. Liu, W.; Salzmann, M.; Fua, P. Context-aware crowd counting. In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5099–5108.
31. Jiang, X.; Xiao, Z.; Zhang, B.; Zhen, X.; Cao, X.; Doermann, D.; Shao, L. Crowd Counting and Density Estimation by Trellis Encoder-Decoder Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
32. Sindagi, V.; Patel, V. Multi-Level Bottom-Top and Top-Bottom Feature Fusion for Crowd Counting. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
33. Mehta, S.; Rastegari, M.; Caspi, A.; Shapiro, L.; Hajishirzi, H. ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
35. Xu, C.; Qiu, K.; Fu, J.; Bai, S.; Xu, Y.; Bai, X. Learn to scale: Generating multipolar normalized density maps for crowd counting. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 8382–8390.
36. Xiong, H.; Lu, H.; Liu, C.; Liu, L.; Cao, Z.; Shen, C. From open set to closed set: Counting objects by spatial divide-and-conquer. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 8362–8371.
37. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
38. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.J. Pytorch: Tensors and dynamic Neural Networks in Python with Strong Gpu Acceleration. 2017. Available online: https://gitee.com/lmy0217/pytorch (accessed on 15 March 2021).
39. Idrees, H.; Tayyab, M.; Athrey, K.; Zhang, D.; Al-Maadeed, S.; Rajpoot, N.; Shah, M. Composition loss for counting, density map estimation and localization in dense crowds. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 532–546.
40. Ieracitano, C.; Paviglianiti, A.; Campolo, M.; Hussain, A.; Pasero, E.; Morabito, F.C. A Novel Automatic Classification System Based on Hybrid Unsupervised and Supervised Machine Learning for Electrospun Nanofibers. *IEEE/CAA J. Autom. Sinica* **2021**, *8*, 68–80. [CrossRef]
41. Chauhan, G.S.; Meena, Y.K.; Gopalani, D.; Nahta, R. A two-step hybrid unsupervised model with attention mechanism for aspect extraction. *Expert Syst. Appl.* **2020**, *161*, 113673. [CrossRef]

42. Zhang, A.; Shen, J.; Xiao, Z.; Zhu, F.; Zhen, X.; Cao, X.; Shao, L. Relational attention network for crowd counting. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6788–6797.
43. Sam, D.B.; Surya, S.; Babu, R.V. Switching convolutional neural network for crowd counting. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4031–4039.
44. Liu, X.; Van De Weijer, J.; Bagdanov, A.D. Leveraging unlabeled data for crowd counting by learning to rank. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7661–7669.
45. Babu Sam, D.; Sajjan, N.N.; Venkatesh Babu, R.; Srinivasan, M. Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Munich, Germany, 8–14 September 2018; pp. 3618–3626.
46. Sindagi, V.A.; Patel, V.M. Generating high-quality crowd density maps using contextual pyramid cnns. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1861–1870.
47. Ranjan, V.; Le, H.; Hoai, M. Iterative crowd counting. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 270–285.