




Article

Crowd Counting Guided by Attention Network

Pei Nie , Cien Fan , Lian Zou, Liqiong Chen  and Xiaopeng Li

School of Electronic Information, Wuhan University, Wuhan 430072, China; niepei@whu.edu.cn (P.N.); zoulian@whu.edu.cn (L.Z.); liqiongchen@whu.edu.cn (L.C.); xiaopengli2014@whu.edu.cn (X.L.)

* Correspondence: fce@whu.edu.cn

Received: 20 September 2020; Accepted: 30 November 2020; Published: 4 December 2020



Abstract: Crowd counting is not simply a matter of counting the numbers of people, but also requires that one obtains people's spatial distribution in a picture. It is still a challenging task for crowded scenes, occlusion, and scale variation. This paper proposes a global and local attention network (GLANet) for efficient crowd counting, which applies an attention mechanism to enhance the features. Firstly, the feature extractor module (FEM) uses the pertained VGG-16 to parse out a simple feature map. Secondly, the global and local attention module (GLAM) effectively captures the local and global attention information to enhance features. Thirdly, the feature fusing module (FFM) applies a series of convolutions to fuse various features, and generate density maps. Finally, we conduct some experiments on a mainstream dataset and compare them with state-of-the-art methods' performances.

Keywords: crowd counting; attention mechanism; global and local attention

1. Introduction

The purpose of crowd counting is to obtain the number of individuals and crowd distribution in a specific scene. Crowd counting has complete video surveillance applications, security monitoring, urban planning, behaviour analysis, and so on. The crowd counting methods are mainly based on detection, regression, and density map. However, it is still a highly challenging task due to occlusion, low quality, resolution, perspective distortion, and scale variations of objects.

With the development of CNNs, various crowd counting method-based CNNs have been proposed in response to this challenging situation. Additionally most of the current work uses Visual Geometry Group (VGG) [1] as a backbone. The receptive field does not change everywhere by using convolution and pooling operations with the same kernel size on the whole image. In the past, this has been solved by combining density maps extracted from image patches of different resolutions or fusing feature maps obtained using convolution filters of various sizes. However, these methods [2–5], by fusing features at all scales, ignore the fact that the scale varies continuously across the image. Later, these methods were proposed using classifiers to predict the size of each patch's receptive domains, which made the network train complicated and not end-to-end. Recent works in crowd counting have been applying attention mechanisms [6,7] to improve network performance and employ perspective maps to guide the accurate estimation of density maps [8–10]. Additionally, most methods [2,4,11] are optimized by comparing the Euclidean distance between the model estimation and the target density map. However, this ignores the connection between pixels and makes the distribution of the crowd blurred.

In recent years, attention models have shown great success in various computer vision tasks [12,13]. Instead of extracting features from the entire image, the attention mechanism allows models to focus on the most relevant features as needed. In this paper, we propose a lightweight attention network to alleviate the effects of various noises in the input, ignore irrelevant information,

and enhance the salient feature extraction for accurate density map estimation. Besides, we combine mean structural similarity [14] loss and Euclidean loss to exploit the local correlation in density maps. The main contributions of this paper are summarized as follows:

- The proposed GLANet generates low-quality density by enhancing different spatial semantic features using multi-column attention mechanisms.
- GLANet utilizes the mean structural similarity to obtain connections between different pixels and the local correlation in density maps.

2. Related Work

The previous literature on crowd counting problems can be divided into three categories according to different methods: detection-based, regression-based and density-based. They can solve or avoid phenomena such as uneven population distribution and overlapping goals, as shown in Table 1.

Table 1. Comparison of the three methods. The \checkmark and \times respectively represent the ability and inability to solve or avoid the phenomenon.

	Detection-Base	Regression-Base	Density-Base
distribution	\checkmark	\times	\checkmark
overlap	\times	\checkmark	\checkmark

2.1. Detection-Based

Detection-based counting methods have deployed a detector to traverse the image to locate and count targets along the way [15–17]. Therefore, people choose to use the more advanced detection to crowd counting. For example, Ref. [15] proposed using head and shoulder detections for crowd counting, and [6] trained a Faster R-CNN [18] for crowd counting by manually annotating the bounding boxes on partial of SHB [2]. However, in crowded scenes, the head sizes can be extremely small, and bounding box annotations is very difficult, which restricts the exploration of detection-based approaches for crowd counting. Besides, the tiny objects have been difficult to handle effectively in previous object detection methods, yet, these objects are very common in crowd counting.

2.2. Regression-Based

Regression-based methods [19–23] directly predict count value of images since it is learnable to map between image features and the crowd count. So regression-based approaches can remedy the occlusion problems, which are difficult for detection-based methods, bypassing explicit detection tasks. More specifically, Ref. [22] has introduced a Bayesian model for discrete regression, which is suitable for crowd counting. However, the regression function is between image features extracted globally from the entire image space, and the total people count in that image, and the spatial information is lost. Although these regression-based methods can accurately estimate the number of people in the picture, they cannot reflect the population's distribution, and most ignore the spatial distribution in the crowd images.

2.3. Density-Based

Density-based approaches effectively count the targets in crowd scenes while maintaining the spatial distribution of the crowd compared with regression-based approaches. In an object density map, the integral over any sub-region is the number of objects within the corresponding region in the image. Therefore, most of the recent methods of crowd counting are based on the density map. Early methods [2,4,8,10,11] advocate a multi-column convolutional architecture to learn the features of different scales by different columns network. For example, Ref. [2] has proposed a three-column network which employs different filters on separate columns to obtain the various scale features,

Ref. [11] adds a classifier to distinguish which column of the picture to use, and [4] has introduced a contextual pyramid CNN that utilizes various estimators to capture both global and local contextual information, which is integrated with high-dimensional feature maps extracted from a multi-column CNN by a Fusion-CNN. Although they have achieved some effect, conflicts caused by optimization between different columns cause training difficulties, and information redundancy between different columns, causing overfitting. Some single column [3,9,24–26] networks are also proposed and attain good performance. For instance, Ref. [25] combines a VGG network with dilated convolution layers to generate a density map, and [3] employs the Inception module to extract features, and deconvolution operator high-resolution density maps.

Due to the perspective distortion, Ref. [27] introduces the additional branch to predict the perspective map, which is used to fuse the multi-resolution density maps. Additionally, as the attention mechanism [28–31] grows, many crowd counting methods [6,32,33] have added it to reduce background interference. Recently, some methods have integrated supervision [23,34] and multi-task [6,35–38] into the crowd counting, and have made considerable progress.

3. Our Approach

3.1. Architecture

The architecture of the proposed network is illustrated in Figure 1, which mainly consists of three components: Feature Extractor Module (FEM), Global and Local Attention Module (GLAM), and Feature Fusing Module (FFM).

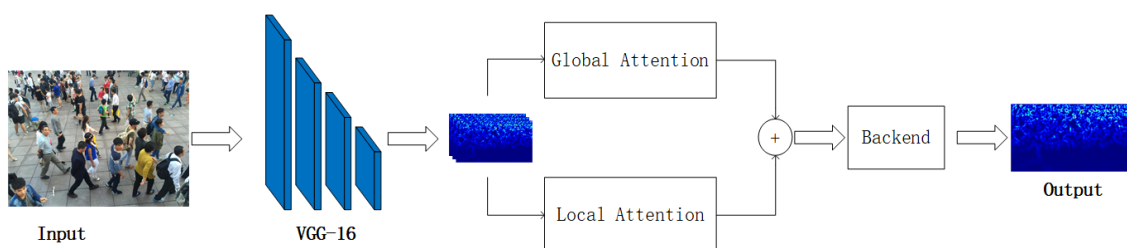


Figure 1. The architecture of the proposed GLANet.

3.1.1. Feature Extractor Module

Following the common practices [9,23,25,33], we chose VGG-16 [1] except for the fully connected layers and two max-pooling layers as the FEM since it has firm transfer learning ability, and its architecture is flexible. Given image I as input, the output I_f produced by the feature extractor can be represented by the following mapping:

$$I_f = F_{VGG}(I) \quad (1)$$

The main difficulty in crowd counting is that the background and the scale of the object have variations in real-world scenes. To apply deep learning for such a situation, a sufficiently large training dataset is required. However, there are few datasets of crowd counting, and the sizes of most datasets are small. As a result, we choose the first ten convolutional layers of a pre-trained VGG-16 with ReLU as FEM.

3.1.2. Global and Local Attention Module

The channel attention block was first introduced as a squeeze and excitation(SE) block in [29] to exploit the inter-channel relationship of features. It utilized the global average pooling to determine the spatial dependency and made specific channel descriptors to emphasize proper channels and recalculate the features map. Additionally, CBAM [28] introduced the spatial attention module to

focus on “where” is an informative part. It applied pooling operations along the channel axis is shown to be effective in highlighting informative regions.

In regular attention operations, spatial attention is separate from channel attention and used separately. Our GLANet introduces an attention mechanism, combining a spatial attention mechanism and channel attention mechanism [28]. As shown in Figure 2, given an intermediate feature map I_f as input, our attention module drives a 3D attention map M , and is summarized as:

$$M = I_f - P_{ave}(I_f) \quad (2)$$

$$I_a = M(I_f) \otimes I_f \quad (3)$$

where \otimes denotes element-wise multiplication. Compared with channel attention, we use global average pooling and upsampling operation to obtain each channel’s background information, instead of the inter-channel relationship. We get the foreground information and we need to pay attention to it by subtracting the feature map’s background information. In crowd counting datasets, there are many unimportant background data that interfere with detection. Additionally, the global average pooling operation will bring a lot of background information. If we use global average pooling to obtain spatial semantic information, it will attain a lot of information that we are not interested in. So, we subtract the features from global average pooling and the input features to eliminate the background information brought by global average pooling and strengthen the original features.

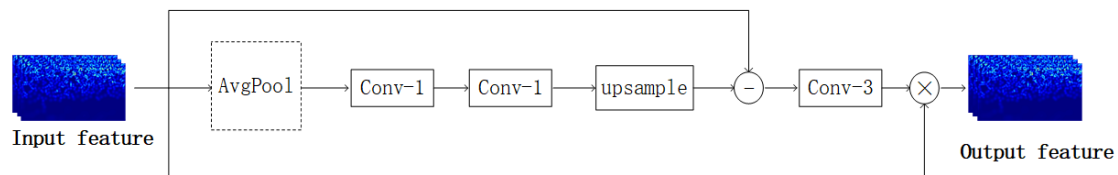


Figure 2. The architecture of the proposed attention module. The parameters of convolutional layers are denoted as Conv-(kernel size). The “AvgPool” refers to the global average pooling operation, and the “upsample” is the upsampling operation.

As is shown in Figure 1, we use a set of attention mechanisms, called global and local attention, in the form of a pyramid to deal with the uneven distribution of people on the picture, and at the same time to make us more aware of crowded areas. In contrast to the pyramid pool network structure [39], it uses global average pooling operators of different sizes to obtain global and local semantic information, which is then concatenated directly with input features. Our attention module uses global average pooling operators of different sizes to get background information of various regions. Additionally firstly, we use a 1×1 convolution to delete some irrelevant information. The attention module is shown in Figure 2.

3.1.3. Feature Fusing Module

To fuse the refined feature maps, we use a series of convolution operators to fuse all feature maps for generating the density map. The structure of FFM is made up of Conv(512,3)—Conv(512,3)—Conv(512,3)—Conv(256,3)—Conv(128,3)—Conv(64,3)—Conv(1,3), and ReLU follows every convolutional layer.

3.2. Density Map Generation

We apply a density map as ground truth to optimize and validate the network. Following the procedure of density map generation in [3], one object at pixel x_i can be represented by a delta

function $\delta(x - x_i)$. As a result, given an image with instances annotated, the ground truth can be represented as:

$$H(x) = \sum_{i=1}^N \delta(x - x_i) \quad (4)$$

In order to generate the density map $F(x)$, the ground truth $H(x)$ is convoluted with a Gaussian kernel, and it can be defined as follows:

$$F(x) = \sum_{i=1}^N \delta(x - x_i) * G_{\sigma_i}(x_i) \quad (5)$$

where σ_i is the standard deviation at the pixel x_i and, empirically, we adopt the fixed kernel with $\sigma = 15$ for all the experiments. The integral over any sub-region is the number of objects within the image's corresponding region in a density map.

3.3. Loss Function

The loss comprises two parts: the pixel-wise Euclidean distance and structural similarity index (SSIM) [14] loss. The proposed framework's training is to minimize a weighted combination loss function concerning the parameters Θ . The final loss function is represented as:

$$L(I; \Theta) = L_2 + \alpha_i L_s \quad (6)$$

where L_2 is the Euclidean distance, L_s is the SSIM loss, and α_i is the weight to balance Euclidean distance and SSIM Loss. In our experiments, the α_i is 100. We use Stochastic Gradient Descent (SGD) with a batch size of 1 for various datasets and a fixed learning rate at 10^{-7} to minimize the loss while the momentum parameter is set as 0.9.

3.3.1. Euclidean Distance

The crowd counting methods based on density map are mostly a regression task, which usually adopts Euclidean distance as a loss function to measure the difference between the estimated density map and ground truth. In symbols, it is defined as:

$$L_2 = \frac{1}{2N} \sum_{i=1}^N \left\| \hat{I}_i - F(I_i) \right\|^2 \quad (7)$$

where \hat{I}_i is the ground truth corresponding to the input image I_i , and $F(I_i)$ is the output density map.

3.3.2. SSIM Loss

The Euclidean distance neglects the local coherence and spatial correlation in density maps. So, we use SSIM loss to enforce the local structural similarity between estimated density map and ground truth. SSIM index is usually used in image quality assessment and computes the similarity between two images from three local statistics—i.e., mean, variance and covariance. Following [14], we use an 11×11 normalized Gaussian kernel with a standard deviation of 1.5 to estimate local statistics. The weight function is defined by $W = \{W(p) \mid p \in P, P = \{(-5, -5), (-4, -4), \dots, (4, 4), (5, 5)\}\}$, where p is offset from the center and P contains all positions of the kernel. For each pixel x on the estimated density map I and the corresponding ground truth \hat{I} , the local statistics are computed by:

$$\mu_I(x) = \sum_{p \in P} W(p) \cdot I(x + p) \quad (8)$$

$$\sigma_I^2(x) = \sum_{p \in P} W(p) \cdot [I(x + p) - \mu_I(x)]^2 \quad (9)$$

$$\sigma_{I\hat{I}}(x) = \sum_{n=P} W(p) \cdot [I(x+p) - \mu_I(x)] \cdot [\hat{I}(x+p) - \mu_{\hat{I}}(x)] \quad (10)$$

where μ_I and σ_I^2 are represented as the local mean and variance estimation of I , $\sigma_{I\hat{I}}$ is local covariance estimation. The SSIM index can be calculated by the following:

$$\text{SSIM} = \frac{(2\mu_I\mu_{\hat{I}} + C_1)(2\sigma_{I\hat{I}} + C_2)}{(\mu_I^2 + \mu_{\hat{I}}^2 + C_1)(\sigma_I^2 + \sigma_{\hat{I}}^2 + C_2)} \quad (11)$$

where C_1 and C_2 are small constants to avoid the occurrence of zero. The SSIM loss is defined as:

$$L_s = 1 - \frac{1}{N} \sum_x \text{SSIM}(x) \quad (12)$$

where N is the number of pixels in density maps.

4. Experiments

A pre-trained VGG-16 initializes the FEM parameters with ReLU, Xavier randomly initializes others with a mean zero and a standard deviation of 0.01. In this section, we introduce datasets and experiment details. Experimental results are compared with the current state-of-the-art methods in Table 2, and the result performance is shown in Figure 3.

Table 2. Comparison with state-of-the-art methods on proposed datasets.

	ShanghaiTech Part A		ShanghaiTech Part B		UCF_CC_50		UCF-QNRF	
Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCNN [2]	110.2	173.2	26.4	41.3	377.6	509.1	277	426
Switch-CNN [11]	90.4	135	21.6	33.4	318.1	439.2	228	445
CP-CNN [4]	73.6	106.4	20.1	30.1	295.8	320.9	295.8	320
SANet [3]	67	104.5	8.4	13.6	258.4	334.9	-	-
CSRNet [25]	68.2	115	10.6	115	266.1	397.5	148	234
DadNet [33]	64.2	99.9	8.8	13.5	285.5	389.7	-	-
D-ConvNet [40]	73.5	112.3	18.7	26	288.4	404.7	-	-
Ours	63.9	104.2	8.3	13	254.6	330	123.8	243

4.1. Evaluation Metrics

The mean absolute error (MAE) and root mean squared error (RMSE) are commonly used to evaluate different methods' performances in previous works. They are defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i^t - y_i| \quad (13)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i^t - y_i)^2} \quad (14)$$

where N is the size of testing images, y_i , and y_i^t are the predicted count and the ground truth for the i -th test image, respectively. Generally speaking, the accuracy of the estimation is shown by the MAE, and the RMSE reflects the robustness of estimation.

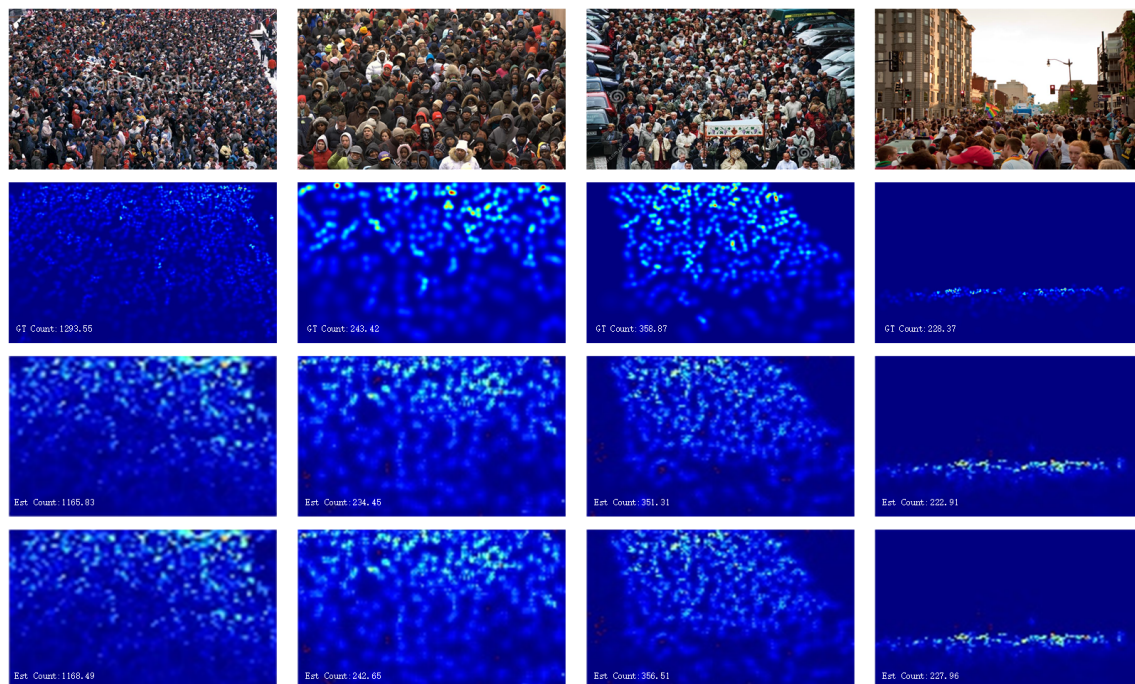


Figure 3. Visualization of density map density maps of three examples for crowd counting. The first row is the input sample image, the second row is the ground truth, the third row is the result of training the GLANet network using Euclidean loss, and the fourth row is the result of training using Euclidean and SSIM loss.

4.2. Dataset

4.2.1. ShanghaiTech

This dataset is made up of 1198 annotated images with a total of 330,165 people head annotated. It is divided into two parts, A and B. The images of part A are randomly collected from the Internet, which consists of 300 training images of different resolutions and 182 testing images, and the scene is mostly highly congested. The average counts are 501.4. Part B images are captured from a relatively sparse crowd in the street, which contain 400 training images and 316 testing images with the same resolutions (768×1024), and the average numbers are 126. During the training, we randomly crop image patches of $1/4$ the original image's size at different locations. These patches are further mirrored to four times the training set.

4.2.2. UCF QNRF

This dataset is the largest crowd dataset which contains 1535 high-resolution images, with 1201 training images and 334 testing images. It annotates approximately 1.25 million people, and the mean counts are 815.4. We crop image patches of $1/9$ the original image's size during the training at different locations, with four patches per original image.

4.2.3. UCF-CC 50

This dataset only has 50 images, but it is annotated with 63,974 in total. It is a challenging task for estimating on this dataset, owing to the number of images and considerable variation in crowd counts, which range from 94 to 4543. We follow the standard protocol and use five-fold cross-validation [41] to evaluate the performance of the proposed method. Ground truth density maps are generated with a fixed spread Gaussian kernel.

4.3. Ablation Study on ShanghaiTech Part A

In order to analyze the effectiveness of our GLA module, we conducted an ablation study on the ShanghaiTech Part A [2] dataset. We removed our Global and Local Attention Module consisting of feature extractor module and feature fusing module. Then we trained it using the same parameters with the fixed learning rate at 10^{-7} . As shown in Table 3, only using the feature extractor module and feature fusing module we obtained the MAE of 68.2 and RMSE of 115 while collecting the MAE of 63.9 and the RMSE of 104.2 to add the Global and Local Attention Module. This achieved 6.3% lower MAE and 9.3% lower RMSE, which demonstrates that the Global and Local Attention Module can significantly decrease the error of estimated crowd count in congested scenes with varied scales.

Table 3. The result of ablation study.

	MAE	MSE
FEM + FFM	69.7	119
Ours	63.9	104.2

5. Conclusions

In this work, we proposed a novel global and local attention network (GLANet) for density maps generation and accurate crowd count estimation. We design an attention mechanism drawing on the spatial attention mechanism and channel attention mechanism to handle the scale variation while keeping the overhead small. Additionally, to exploit the local correlation of density maps, we used the SSIM loss to enforce the local structural similarity between density maps. Extensive experiments show that our method achieves the superior performance on four major crowd counting benchmarks to state-of-the-art methods.

Author Contributions: Methodology, P.N.; Program, P.N. and C.F.; Formal Analysis, P.N.; Data Curation, L.Z.; Supervision, L.C.; Resources, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

References

1. 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](#)]
2. 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.
3. 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.
4. 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.
5. Zhao, M.; Zhang, J.; Zhang, C.; Zhang, W. Leveraging heterogeneous auxiliary tasks to assist crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12736–12745.
6. Liu, J.; Gao, C.; Meng, D.; Hauptmann, A.G. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5197–5206.
7. Zhang, Y.; Zhou, C.; Chang, F.; Kot, A.C.; Zhang, W. Attention to head locations for crowd counting. In Proceedings of the International Conference on Image and Graphics. Springer, Beijing, China, 23–25 August 2019; pp. 727–737.

8. Onoro-Rubio, D.; López-Sastre, R.J. Towards perspective-free object counting with deep learning. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 615–629.
9. Shi, M.; Yang, Z.; Xu, C.; Chen, Q. Revisiting perspective information for efficient crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7279–7288.
10. 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.
11. 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.
12. Abdolrashidi, A.; Minaei, M.; Azimi, E.; Minaee, S. Age and Gender Prediction From Face Images Using Attentional Convolutional Network. *arXiv* **2020**, arXiv:2010.03791.
13. Li, L.; Tang, S.; Deng, L.; Zhang, Y.; Tian, Q. Image caption with global-local attention. In Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4133–4139.
14. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]
15. 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 IEEE 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–4.
16. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1627–1645. [[CrossRef](#)] [[PubMed](#)]
17. Wang, M.; Wang, X. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In Proceedings of the IEEE CVPR, Providence, RI, USA, 20–25 June 2011; pp. 3401–3408.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; IEEE: Montreal, QC, Canada, 2015; pp. 91–99.
19. Paragios, N.; Ramesh, V. A MRF-based approach for real-time subway monitoring. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kauai, HI, USA, 8–14 December 2001; Volume 1, p. I.
20. Regazzoni, C.S.; Tesei, A. Distributed data fusion for real-time crowding estimation. *Signal Process.* **1996**, *53*, 47–63. [[CrossRef](#)]
21. Chan, A.B.; Liang, Z.S.J.; Vasconcelos, N. Privacy preserving crowd monitoring: Counting people without people models or tracking. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–7.
22. Chan, A.B.; Vasconcelos, N. Counting people with low-level features and Bayesian regression. *IEEE Trans. Image Process.* **2011**, *21*, 2160–2177. [[CrossRef](#)] [[PubMed](#)]
23. 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.
24. Liu, L.; Wang, H.; Li, G.; Ouyang, W.; Lin, L. Crowd counting using deep recurrent spatial-aware network. *arXiv* **2018**, arXiv:1807.00601.
25. 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; pp. 1091–1100.
26. Chen, X.; Bin, Y.; Sang, N.; Gao, C. Scale pyramid network for crowd counting. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1941–1950.
27. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
28. Woo, S.; Park, J.; Lee, J.Y.; So Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.

29. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
30. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*; NIPS'14: Montreal, QC, Canada, 2014; pp. 2204–2212.
31. Luong, M.T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. *arXiv* **2015**, arXiv:1508.04025.
32. Hou, Y.; Li, C.; Yang, F.; Ma, C.; Zhu, L.; Li, Y.; Jia, H.; Xie, X. BBA-NET: A Bi-Branch Attention Network For Crowd Counting. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 4072–4076.
33. Guo, D.; Li, K.; Zha, Z.J.; Wang, M. Dadnet: Dilated-attention-deformable convnet for crowd counting. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1823–1832.
34. Von Borstel, M.; Kandemir, M.; Schmidt, P.; Rao, M.K.; Rajamani, K.; Hamprecht, F.A. Gaussian process density counting from weak supervision. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 365–380.
35. Sindagi, V.A.; Patel, V.M. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; pp. 1–6.
36. Shi, Z.; Mettes, P.; Snoek, C.G. Counting with focus for free. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 4200–4209.
37. Jiang, S.; Lu, X.; Lei, Y.; Liu, L. Mask-aware networks for crowd counting. *IEEE Trans. Circuits Syst. Video Technol.* **2019**. [[CrossRef](#)]
38. Gao, J.; Wang, Q.; Li, X. PCC Net: Perspective Crowd Counting via Spatial Convolutional Network. *IEEE Trans. Circuits Syst. Video Technol.* **2019**. [[CrossRef](#)]
39. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
40. Shi, Z.; Zhang, L.; Liu, Y.; Cao, X.; Ye, Y.; Cheng, M.-M.; Zheng, G. Crowd counting with deep negative correlation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5382–5390.
41. Idrees, H.; Saleemi, I.; Seibert, C.; Shah, M. Multi-source multi-scale counting in extremely dense crowd images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2547–2554.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).