Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 2, No. 1, 2022

Hybrid Attention Fusion in Dense Crowd Counting

Suyu Han*

College of Computer Science & Technology, Qingdao University, Qingdao 266071, Shandong, China
* **Corresponding author**: Email: mail.hsy@qq.com.

Abstract: One of appealing approaches to guiding deep parameter optimization, is attentional supervision, which inspires intelligence in complex networks at a fraction of the cost, but there is still room for improvement. First, the real dense scene with varying scales and uneven density distribution of human heads, the density map cannot be clearly expressed. Second, the heavily occluded areas are extremely similar to the complex background, which further aggravates the counting error. Therefore, we propose a dual-track attention network that distinguishes between global and local information, which is responsible for the target overlap and background confusion problems, respectively, and finally converges and normalizes with the feature map to transform the multi-channel attention map into a single-channel density map. Meanwhile the heterogeneous pyramid design alleviates the distress of scale variation and density dissimilarity. Experiments on several official datasets prove the effectiveness of the scheme to enhance key information and overcome confounding factors.

Keywords: Crowd counting; Attention fusion; SoftMax algorithm; Density map.

1. Introduction

Dense crowd counting is defined as estimating the number of people in an image or video clip, generally using heads as the counting unit. In the density map estimation strategy, each pixel point represents the probability of this location being the center of the head, thus reducing the counting task to the accumulation of probabilities. However, in real scenarios, a robust counting model requires strong generalization ability to external disturbances such as noisy background, scale variation, mutual occlusion, perspective distortion, etc. Traditional counting networks rely on multi-column architectures to extract features at different scales while focusing on valuable visual information based on attention mechanisms.

1.1. Multi-scale Feature Extraction Strategy

This strategy emphasizes that targets at different scales need to be perceived by perceptual fields of different sizes, which is generally implemented using a multi-column convolutional architecture. MCNN [1] first uses a three-column network architecture to extract multi-scale features to accommodate scale variations due to different camera angles. Inspired by MCNN [1], Switching-CNN [2] retains the multi-column mode, adds a classifier to select the best branch suitable for the current scale, and then adaptively fuse the multi-scale information. Further, CSRNet [3] adapts a dilated convolutional layer to increase the receptive field as an alternative to the pooling operations, but it tends to lead to grid effects, further leading to local information loss. SANet [4] sets inception layout in the encoder to extract multi-scale features and adds transposed convolution in the decoder to generate high-resolution density maps.

1.2. Attention Mechanism Guidance Strategy

Attention mechanism is activated by the sigmoid function, which directs the model to focus on regions where the signal response is obvious and suppresses background noise, thus acting as a top-level supervision. ASNet [5] considers the density of different regions in an image varies greatly, leading to heterogeneous counting performance, and therefore proposes density attention networks to provide multi-scale attention masks for convolutional extraction units. HANet [6] utilizes progressive embedding of scale-context fusion channel attention with spatial attention, without considering that there are differences in the supervised objects of attention in the local and global cases. RANet [7] emphasizes on attention optimisation, using two modules to handle global attention and local attention separately, and then finally fusing them based on the interdependencies between features, but the dependencies are difficult to determine. Recognizing that it is often difficult to generate accurate attention maps directly, CFANet [8] turns to a coarse-to-fine progressive attention mechanism through two branches, the crowd region identifier (CRR) and the density level estimator (DLE).

2. Proposed Method

2.1. The Main Network Structure

This paper aims to establish a crowd counting framework that is suitable for dense scenes. The architecture of the proposed method is illustrated in Figure 1. It includes a primary feature extractor taken from the VGG-16 [9] model as the backbone, two heterogeneous pyramid modules acting on global and local information encoding, respectively, and dual-track stacking to obtain the hybrid features, activated to get attention and then fused with the multi-channel feature map to obtain the final predicted density map.

The local information encoding stage is a shallow network, capable of exploiting more fine-grained feature information and thus filtering complex backgrounds with other non-target entities. A four-branch pyramidal architecture is specifically used, with increasing convolutional kernel size from top to bottom and progressively larger perceptual fields.

The global information encoding stage has a deeper layer of network with enhanced nonlinearity, fuller perceptual field, and richer semantic information. It is used to deal with the dense occlusion of human head and has good learning ability for irregular density distribution as well. To reduce the parameter overhead, a three-way merge is used, and the filter size is also chosen among 1×1, 3×3, and 5×5.

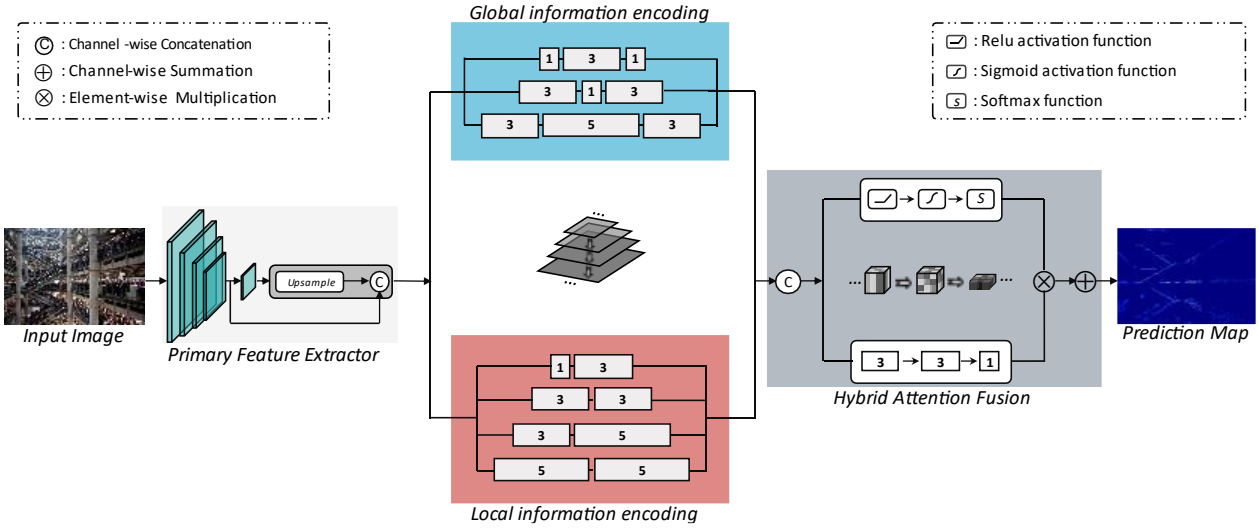

Figure 1. Overall architecture of the proposed network

The dual-track features are merged and split in two again. In the top-side pathway, the probability distribution between [0,1] is generated via ReLU and Sigmoid activation functions in turn, i.e., hybrid attention. And in the bottom-side pathway, two 3×3 convolution kernels are used to reform the feature information first, and then 1×1 convolution is used to adjust it to the exact same size as the hybrid attention, i.e., the multi-channel feature map. For the fusion strategy, we introduce the softmax function, which converts the multi-channel attention map into a single-channel density map, a move that eliminates the need for the network to resort to costly and poorly robust attention labels. In detail, the hybrid attention is normalized by softmax and each pixel can learn the dynamic weight of this location in all channel layers. Then, it is multiplied with the multi-channel feature maps pairwise, and finally, all channels are summed vertically to obtain the final prediction map of fused attention to features, denoted by $F_{pre} \in \mathbb{R}^{1\times H\times W}$.

$$F_{pre_{i,j}} = \sum_{k=1}^{C}\left(SoftMax\left(F_{att}\right)\otimes F_{mul}\right)_{k,i,j} \quad \begin{cases}1\le i\le H\\ 1\le j\le W\end{cases} \tag{1}$$

2.2. Loss Function

In this paper, we choose the mean absolute error (MAE) to measure the pixel-level error values between the final prediction map and the labels, denoted by L_{pre}.

$$L_{pre} = \frac{1}{N}\sum_{i=1}^{N}\left\|P\left(X_i;\Theta\right)-G_i^{GT}\right\|_2^2 \tag{2}$$

Where N is the number of images in a training batch, X_i denotes the current training image, Θ is a set of learnable parameters, so $P(X_i;\Theta)$ represents the prediction map for it, and G_i^{GT} refers to its ground-truth density map.

3. Experiments and Results Analysis

3.1. Experimental Detail

To ensure the experimental authority, four official datasets, ShanghaiTech(A&B) [1], UCF_CC_50 [10], and UCF-QNRF [11], are used in this paper. Among them, the UCF_CC_50 dataset has a limited number of samples, and we follow the official recommendation of 5-fold cross-validation for testing.

We generate training labels by blurring each head annotation with a Gaussian function. In detail, for crowd-sparse datasets, such as ShanghaiTech Part B, we use fixed-size kernels, while for other datasets with denser scenes, geometric adaptive kernel based on the nearest neighbor algorithm is utilized.

Except for Primary Feature Extractor, the parameters of the subsequent layers are randomly initialized by a Gaussian distribution with a mean of 0 and a standard deviation of 0.01. For training details, we choose the Adam optimizer to retrain the model, with an initial learning rate of 1E-4, halved every 100 rounds.

There are two mainstream metrics for evaluating the performance in crowd counting task: mean absolute error (MAE) and mean squared error (MSE). They are defined as follows.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|P_i-G_i\right| \tag{3}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|P_i-G_i\right|^2} \tag{4}$$

3.2. Comparison Experiment

We demonstrate the effectiveness of the proposed method on four official datasets, and the experimental results are shown in Table 1 (The best performance is indicted by bold and the second best is underlined). In the ShanghaiTech Part_A dataset, our MAE is 0.33% ahead of HANet; for the UCF_CC_50 dataset, it outperforms the ASNet result by 4.5%, while the MSE is 2.59% ahead. Also, in the UCF-QNRF dataset, we improve the MSE metric by 1.03%.

To visually compare the effectiveness of the proposed method with RANet, we select the representative samples from each dataset for counting tests, as shown in Figure 2. Further, to observe the overall prediction effect of the two on ShanghaiTech Part_A dataset, we aggregate the PRE-GT information for the entire sample in this dataset and plot it as a scatter diagram with regression lines, the results of which are shown in Figure 3, where the red auxiliary line y=x indicates the ideal case of 100% accuracy in counting. Qualitatively, the closeness of the blue regression line to the auxiliary line y=x is positively correlated with the quality of the prediction; quantitatively, the closer the coefficient of determination R^2 of the regression line is to 1, the lower the overall error fluctuation is.

Table 1. Comparison with different methods on four challenging datasets

Method	Part_A		Part_B		UCF_CC_50		UCF-QNRF	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
BL[12]	62.8	101.8	7.7	12.7	229.3	308.2	88.7	154.8
RANet[7]	59.4	102	7.9	12.9	239.8	319.4	111	190
ASNet[5]	57.78	**90.13**	-	-	174.84	251.63	91.59	159.71
LibraNet[13]	55.9	97.1	7.3	11.3	181.2	262.2	88.1	143.7
AMSNet[14]	56.7	93.4	6.7	**10.2**	208.4	297.3	101.8	163.2
UOT[15]	58.1	95.9	**6.5**	**10.2**	-	-	83.3	142.3
URC[16]	68.2	115	10.6	16	293.99	443.09	128.13	218.05
DKPNet[17]	55.6	91	6.6	10.9	-	-	**81.4**	147.2
HANet[6]	54.9	91.2	6.8	11.5	195.2	268.6	98	179
D2C[18]	59.6	100.7	6.7	10.7	221.5	300.7	84.8	145.6
Ours	**54.72**	93.76	7.12	12.86	**166.97**	**245.12**	83.02	**140.84**

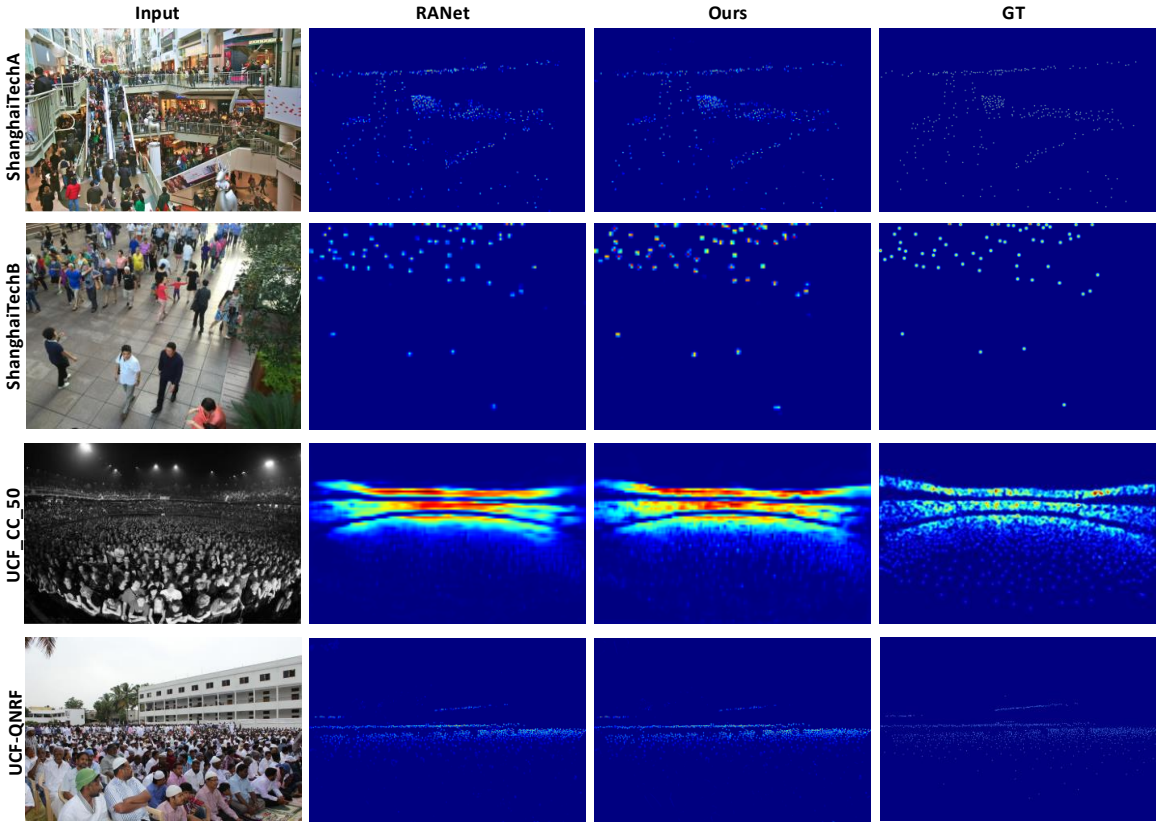

Figure 2. Visual comparison of different methods

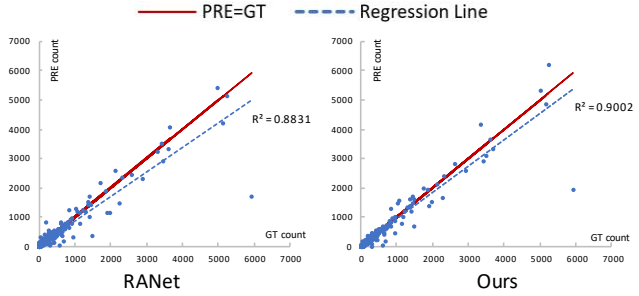

Figure 3. Scatter plot vs. regression line for different methods

4. Conclusion

In this paper, we dissect the current problems faced by dense crowd counting, including noisy background, mutual occlusion and variable scale. A dual-track network is designed, using heterogeneous pyramid modules to obtain global and local features respectively, which are transformed into hybrid attention based on the softmax algorithm and fused with high-resolution feature maps to effectively deal with the problem of responsible target overlap and background confusion, on the other hand, the pyramid paradigm itself has strong learning capability for variable scales. Experiments show that this strategy is effective and has more stable performance.

References

[1] Zhang Y, Zhou D, Chen S, et al. Single-image crowd counting via multi-column convolutional neural network[C] //Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 589-597.

[2] Babu Sam D, Surya S, Venkatesh Babu R. Switching convolutional neural network for crowd counting[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5744-5752.

[3] Li Y, Zhang X, Chen D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1091-1100.

[4] Cao X, Wang Z, Zhao Y, et al. Scale aggregation network for accurate and efficient crowd counting[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 734-750.

[5] Jiang X, Zhang L, Xu M, et al. Attention scaling for crowd counting[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 4706-4715.

[6] Wang F, Sang J, Wu Z, et al. Hybrid attention network based on progressive embedding scale-context for crowd counting[J]. Information Sciences, 2022, 591: 306-318.

[7] Zhang A, Shen J, Xiao Z, et al. Relational attention network for crowd counting[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6788-6797.

[8] Rong L, Li C. Coarse-and fine-grained attention network with background-aware loss for crowd density map estimation[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021: 3675-3684.

[9] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[10] Idrees H, Saleemi I, Seibert C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 2547-2554.

[11] Idrees H, Tayyab M, Athrey K, et al. Composition loss for counting, density map estimation and localization in dense crowds[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 532-546.

[12] Ma Z, Wei X, Hong X, et al. Bayesian loss for crowd count estimation with point supervision[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6142-6151.

[13] Liu L, Lu H, Zou H, et al. Weighing counts: Sequential crowd counting by reinforcement learning[C]//European Conference on Computer Vision. Springer, Cham, 2020: 164-181.

[14] Hu Y, Jiang X, Liu X, et al. Nas-count: Counting-by-density with neural architecture search[C]//European Conference on Computer Vision. Springer, Cham, 2020: 747-766.

[15] Ma Z, Wei X, Hong X, et al. Learning to count via unbalanced optimal transport[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(3): 2319-2327.

[16] Xu Y, Zhong Z, Lian D, et al. Crowd counting with partial annotations in an image[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 15570-15579.

[17] Chen B, Yan Z, Li K, et al. Variational attention: Propagating domain-specific knowledge for multi-domain learning in crowd counting[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 16065-16075.

[18] Cheng J, Xiong H, Cao Z, et al. Decoupled two-stage crowd counting and beyond[J]. IEEE Transactions on Image Processing, 2021, 30: 2862-2875.