ADAPTIVE SCENARIO DISCOVERY FOR CROWD COUNTING

Xingjiao Wu[1], *Yingbin Zheng*[2,3], *Hao Ye*[2,3], *Wenxin Hu*[1*], *Jing Yang*[1], *Liang He*[1]

[1]East China Normal University, Shanghai, China
[2]Shanghai Advanced Research Institute, CAS [3]Videt Tech.
52184506007@stu.ecnu.edu.cn, {yingbin.zheng, hao.ye}@videt.cn, wxhu@cc.ecnu.edu.cn, {jyang, lhe}@cs.ecnu.edu.cn

ABSTRACT

Crowd counting, *i.e.*, estimation number of the pedestrian in crowd images, is emerging as an important research problem with the public security applications. A key component for the crowd counting systems is the construction of counting models which are robust to various scenarios under facts such as camera perspective and physical barriers. In this paper, we present an adaptive scenario discovery framework for crowd counting. The system is structured with two parallel pathways that are trained with different sizes of the receptive field to represent different scales and crowd densities. After ensuring that these components are present in the proper geometric configuration, a third branch is designed to adaptively recalibrate the pathway-wise responses by discovering and modeling the dynamic scenarios implicitly. Our system is able to represent highly variable crowd images and achieves state-of-the-art results in two challenging benchmarks.

Index Terms— Crowd counting, adaptive scenario discovery, convolutional neural network.

1. INTRODUCTION

Counting is the process of estimating the number of a particular object. With the expansion of urban population and the convenience of modern transportation, it is common to have large crowds in specific events or scenarios, and crowd counting from images or videos becomes crucial for applications ranging from traffic control to public safety.

Previous methods of crowd counting may be roughly divided into two categories: detection-based and regression-based. Detection-based methods have been studied with the pedestrian detectors [1, 2]. However, it is challenging for these methods to model a very dense crowd or crowd in a clustered environment. The regression-based approaches are firstly proposed in [3]. With the recent development of the convolutional neural network (CNN), the regression framework by estimation of the density maps has been widely used.

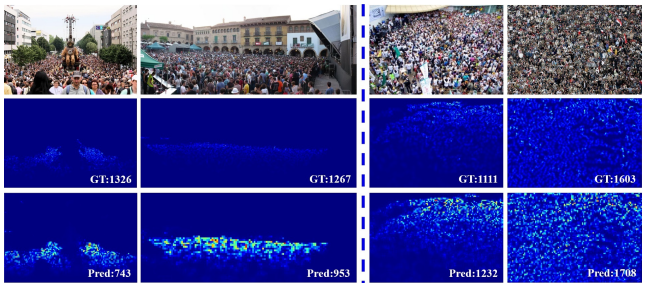


Fig. 1. The crowd images from the ShanghaiTech dataset [5] and their crowd counting prediction by CSRNet [8].

Compared with the system employing a single CNN regressor (*e.g.*, [4]), the networks with multiple columns/branches learn more contextual information and achieve excellent performance [5, 6, 7, 8]. Although different receptive fields are usually applied in multiple branches, it is difficult to represent highly variable crowd images. There still exist gaps between the ground-truth and prediction for some crowd images (some examples are shown in Fig. 1). We also observe that the images under similar scenario seem to have the same prediction pattern: the images with the lower camera viewpoints and more backgrounds usually achieve smaller counting prediction than the ground-truth (Fig. 1-Left), while these with high viewpoint get larger predicted values (Fig. 1-Right).

The central issue addressed in this paper is the following: *Can we design a model to discover the scenarios and modeling the crowd images simultaneously?* One intuitive idea is to add the number of network branch with well-designed convolution filters. The limitations are, the CNN model will be difficult to train with the current crowd counting datasets, and it is also hard to directly define the scenarios. In this paper, we present an adaptive scenario discovery framework for crowd counting. Our network adopts the VGG model [9] as the backbone and is structured with two parallel pathways that are trained with different sizes of the receptive field to serve different scales and crowd densities. We consider the scenario as a linear combinational of two pathway with the discretized weights and design a third adaption branch to learn this scenario aware responses and discover the scenarios implicitly.

*Corresponding author.

This work was supported in part by grants from NSFC (#61602459) and STCSM's program (#18511103105). The computation of this work was performed in the Supercomputer Center of ECNU.

Our contributions are summarized as follows.

- From the perspective of scenario discovery, a novel adaptive framework for crowd counting is proposed. Different from previous multiple columns/branches frameworks, ours has the ability to represent highly variable crowd images with two branches by incorporating the discretized pathway-wise responses.
- We apply our framework to the ShanghaiTech [5] and UCF_CC_50 [3] crowd counting datasets, and find that it outperforms the state-of-the-art approaches.

1.1. Related Work

Numerous efforts have been devoted to the design of crowd counting models. Detail survey of the recent progress can be found in [10]. In this section, we mainly discuss literature on the models with multiple branches representation, which are more related to this work. In [5], Zhang *et al.* proposed the MCNN by using three columns of convolutional neural networks with filters of different sizes. Sam *et al.* [6] proposed the Switching-CNN, which decoupled the three columns into to separate CNN (each trained with a subset of the patches), and a density selector is designed to utilize the structural and functional differences. Several works have studied the context information of the crowd images under multiple branch setting. For instance, Sindagi *et al.* [7] applied local and global context coding to population count density estimation, and Zhang *et al.* [11] proposed a scale-adaptive CNN architecture with a backbone of fixed small receptive fields. Another work related to ours is the CSRNet [8], where convolutional neural networks with dilation operations were employed after the backbone of the pre-trained deep model.

These existing approaches construct density estimation models with multiple branches to represent different receptive fields or scales. Our framework also follows the general process, with the design of one branch representing the dense prediction and another for the relative sparse crowds. However, instead of using the fix branch weights or selecting one explicitly column, we adopt the learning of branch weights. Responses of the dense and sparse pathways are adaptively recalibrated by a third branch, which explicitly models interdependencies between pathways. Moreover, with the discretization of these pathway-wise responses, the crowd scenarios are implicitly discovered and respond to different crowd images in a highly scenario-specific manner. The whole framework can be end-to-end trained, and as will be shown in the experiments, it is more accurate compared to previous approaches.

2. FRAMEWORK

The overall architecture of our framework is illustrated in Fig. 2. We start by introducing the design of adaptive scenario discovery, followed by implementation details of the framework.

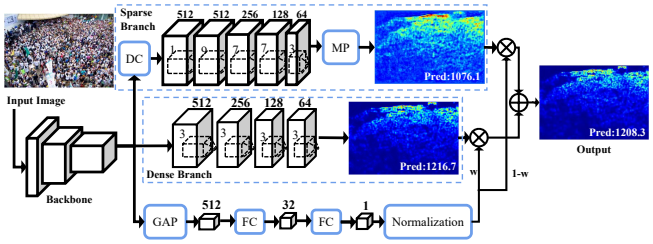


Fig. 2. Network structure of proposed framework. DC, MP, GAP, and FC indicate deconvolution layer, max pooling, global average pooling, and full-connected layer respectively.

2.1. Adaptive Scenario Discovery

The selection of a suitable network structure is important to the success of a crowd counting system. There are generally two categories of networks: either it is with a new design of the structure and learned from scratch (*e.g.*, [5, 12]), or the model is transferred from part of a pre-trained network (*e.g.*, [8, 13]). In this paper, our framework belongs to the second case, by employing the convolutional layers of a VGG-16 model [9] pre-trained from the ImageNet dataset [14] and fine-tuned with the crowd images. We choose this strategy for the outstanding performance of the model in crowd counting as well as other computer vision tasks, and the results in the evaluation also confirm the effects of the pre-trained model.

Our counting network consists of two parallel pathways after the backbone module. The first pathway starts with a deconvolution layer that amplifies the inputs, and then a few convolutional layers with larger receptive fields are used, followed by a 2×2 max pooling. This pathway is designed to model the high congested scenario with *dense* crowd, and the second pathway is for the *sparse* scenario. The convolution filers in this subnet are with a size of 3×3. Note that the concept of dense or sparse is relative and both pathways can output a density map.

There are several approaches to fuse the density maps, and here we would like to use a dynamic weighting strategy. Inspired by the excitation operation in SENet [15], we propose the *adaption* branch. The outputs of the last convolutional layer in the backbone go through a global average pooling and two fully-connected layers and then have an initial response w. We expect w to adaptively recalibrate the weight of the dense and sparse pathways, therefore we normalize it into the interval of [0,1) with the following formula:

$$w^* = \arctan(\text{sigmod}(w)) \times \frac{2}{\pi} \tag{1}$$

Experiments on Section 3.2 will show the effect compared with the single branch or average fusion. However, we find that the convergence speed of this architecture is slow, proba-

bly due to the small size of the crowd counting dataset but the continuous response.

Our solution is to divide the response value into bins, by borrowing the idea from traditional visual features such as color histogram [16], SIFT [17], and HoG [18]. The benefits of discretization are two-folder. First, the model itself is easier to train and converge. Second, similar attributes are significantly observed from the images within the same bin (see Fig. 5), indicating that discretization operation is able to implicitly discovering the dynamic scenarios.

2.2. Implementation Details

Ground Truth Generation. We follow [8] to generate the density maps from ground truth. the density map $F(x)$ is generated with the formula:

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

where x_i is a targeted object in the ground truth δ and $G_{\sigma_i}(\cdot)$ is a Gaussian kernel with standard deviation of σ_i. For the datasets with high congested scene (such as ShanghaiTech Part A [5] and UCF_CC_50 [3]), $F(x)$ is defined as a geometry-adaptive kernel with $\sigma_i = \beta \bar{d}_i$. Here $\bar{d}_i$ is the average distance of k nearest neighbors of targeted object x_i. For low congested scene (*i.e.*, ShanghaiTech Part B [5]), we set $\sigma_i = 15$.

Training Details. We define the loss function as follows:

$$L(\Theta) = \frac{1}{2N} \sum_{i=1}^{N} \|\mathcal{F}(X_i; \Theta) - F(X_i)\|_2^2 \tag{3}$$

where $F(X_i)$ is the ground truth density map of image X_i from Equ. (2) and $\mathcal{F}(X_i; \Theta)$ is the estimated density map of X_i with the parameters Θ learned by the proposed network.

To ensure the spatial feature and the context of the crowd images, we do not extract the image patches for data augmentation. And there is also no additional image copy/conversion enhancement. During training, we employ the stochastic gradient descent (SGD) for its good generalization ability.

3. EVALUATIONS

We conduct the experiments on the ShanghaiTech dataset [5] and the UCF_CC_50 dataset [3]. The ShanghaiTech dataset [5] is divided into Part A and Part B. ShanghaiTech Part A contains 482 crowd images with 300 training images and 182 testing images, and the average number of the pedestrian is 501. ShanghaiTech Part B is with 716 images (400 training and 316 testing). The resolution of the images are fixed with 768×1024 pixels, and the pedestrian number is generally smaller than Part A with an average number of 123. The UCF_CC_50 dataset [3] contains 50 images with high crowd density. The images vary in the number of pedestrians, with a range of 94 to 4,543. For both datasets, we follow the standard experimental protocols, and mean absolute error (MAE) and mean squared error (MSE) is reported as the evaluation metric. We implement our framework based on PyTorch [24].

Table 1. Comparison with the state-of-the-arts on the benchmarks. Part A and Part B indicate ShanghaiTech Part A and Part B, respectively.

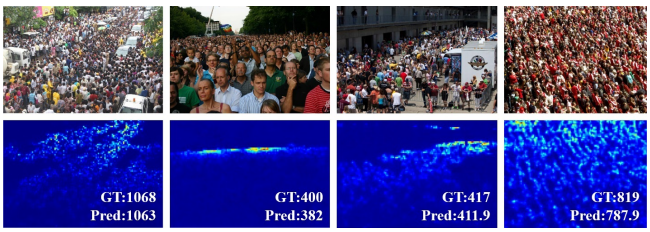
Method	Part A		Part B		UCF_CC_50	
	MAE	MSE	MAE	MSE	MAE	MSE
Zhang et al. [19]	181.8	277.7	32.0	49.8	467.0	498.5
MCNN [5]	110.2	173.2	26.4	41.3	377.6	509.1
Cascaded-MTL [20]	101.3	152.4	20.0	31.1	322.8	397.9
Switching-CNN [6]	90.4	135.0	21.6	33.4	318.1	439.2
DAN [21]	88.5	147.6	17.6	26.8	234.5	289.6
CP-CNN [7]	73.6	106.4	20.1	30.1	295.8	320.9
Huang *et al.* [22]	-	-	20.2	35.6	409.5	563.7
D-ConvNet [13]	73.5	112.3	18.7	26.0	288.4	404.7
ACSCP [12]	75.7	102.7	17.2	27.4	291.0	404.6
DecideNet [23]	-	-	20.8	29.4	-	-
SaCNN [11]	86.8	139.2	16.2	25.8	314.9	424.8
CSRNet [8]	68.2	115.0	10.6	16.0	266.1	397.5
ASD [ours]	**65.6**	**98.0**	**8.5**	**13.7**	**196.2**	**270.9**

3.1. Results and Comparison

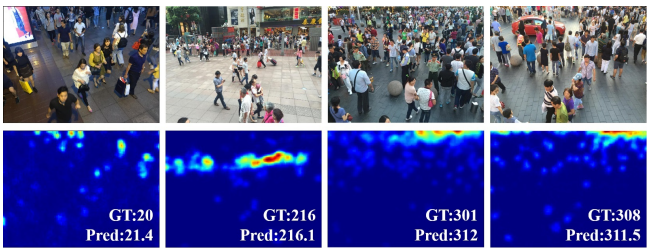
We first evaluate the overall results of our proposed framework. We compare our framework with several state-of-the-art approaches, including the multi-column CNN with different receptive fields [5], the Switching-CNN that leverages variation of crowd density [6], and a very recent dilated convolution based model CSRNet [8]. The number of grouped scenario is 15, and the effect of the parameters will be evaluated in the next subsection. We denote our approach as *ASD* (*A*daptive *S*cenario *D*iscovery) in the following comparisons.

ShanghaiTech. Table 1 summarizes the MAE and MSE of previous approaches and ours in the datasets. On Part A of ShanghaiTech, we achieve a significant overall improvement of 24.8 of absolute MAE value over Switching-CNN [6] and 2.6 of MAE over the state-of-the-art CSRNet [8]. On Part B, our ASD framework also achieves the best MAE 8.5 and MSE 13.7 compared to the state-of-the-art. Fig. 3(a) and (b) illustrate the density maps and the prediction results of some crowd images from both parts respectively.

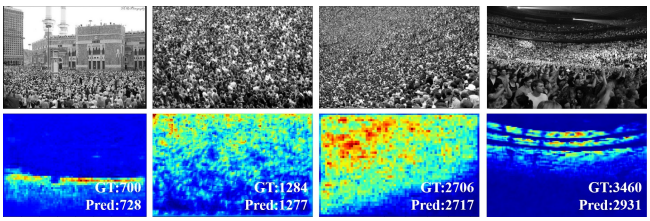
UCF_CC_50. We now report results on the UCF_CC_50 dataset, as summarized in Table 1 and shown in Fig. 3(c). Similar to the experiments on ShanghaiTech, the ASD framework shows better results than the other approaches, and we improve on the previously reported state-of-the-art results by 26.3% for the MAE metric and 31.8% for the MSE, which indicates the low variance of our prediction across the high crowd density images.



(a) ShanghaiTech Part A



(b) ShanghaiTech Part B



(c) UCF_CC_50

Fig. 3. Qualitative results on the benchmarks.

3.2. Ablation Study

In this part, we evaluate a few parameters and an alternative implementation for the proposed framework. We report results on the ShanghaiTech Part A.

Network Architecture. We first evaluate the effect of the two parallel pathways over the whole framework. Fig. 4-Left gives the comparison with different network architecture, including the single pathway and the fusion of them. With the fusion of a fixed pathway weight, the result is 74.1 of MAE and 114.0 of MSE, which is not higher than results by the single pathway. We observe significant performance gains when adding the dynamic pathway-wise responses and the discretization.

The Effect of Scenario Discovery. Recall that the discretization on the adaption branch is applied to discover the dynamic scenarios implicitly; here we consider the different choice of parameters. The output response after the operation of Equ. (1) fall in the interval (0,1), and is divided into 2,10,100, and 1000 bins. Note that only a proportion of bins are with images after model training due to the size of the dataset, therefore the number of scenarios is usually smaller than that of the bin. Discretization with 2 bins can be considered as a simplified version of Switching-CNN [6], and our learning strategy

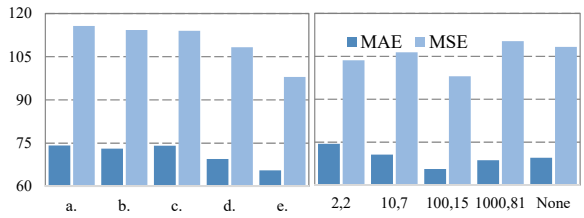


Fig. 4. Left: the effect of varying network architecture, a. sparse pathway only; b. dense pathway only; c. fusion of the two pathways with the same weight; d. learned weight without discretization; e. proposed approach. Right: the effect of scenario discovery w.r.t the number of discretization bins and grouped scenarios ("None" indicates the result without discretization).

Fig. 5. Images of four sample scenarios grouped by adaptive scenario discovery. A various of differences between each two scenarios, such as crowd density, ratio of background, and viewpoints, can be visibly from the images.

still achieves lower MAE (74.4 vs. 90.4). Without the discretization, we obtain the MAE of 69.4, which is not as good as the scenario discovery with 15 and 81 scenarios (MAE of 65.6 and 68.7, respectively). Fig. 5 shows some crowd images from different scenarios.

4. CONCLUSIONS

In this paper, we have presented a novel architecture for high-density population counting. Our approach focuses on the implicit discovery and dynamic modeling of scenarios. In addition, we have reformulated the crowd counting problem as a scenario classification problem such that the semantic scenario models into a combined prediction sub-tasks. The adaptive scenario discovery is built to obtain two weights of different sizes through the parallel perception path for dynamic fusion. Our proposed framework achieves state-of-the-art performance on two popular crowd counting datasets.

5. REFERENCES

[1] M. Wang and X. Wang, "Automatic adaptation of a generic pedestrian detector to a specific traffic scene," in *CVPR*, 2011, pp. 3401–3408.

[2] R. Stewart, M. Andriluka, and A. Y. Ng, "End-to-End people detection in crowded scenes," in *CVPR*, 2016, pp. 2325–2333.

[3] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *CVPR*, 2013, pp. 2547–2554.

[4] D. Oñoro Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *ECCV*, 2016, pp. 615–629.

[5] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *CVPR*, 2016, pp. 589–597.

[6] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *CVPR*, 2017, pp. 5744–5752.

[7] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid cnns," in *ICCV*, 2017, pp. 1879–1888.

[8] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *CVPR*, 2018, pp. 1091–1100.

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

[10] V. A Sindagi and V. M Patel, "A survey of recent advances in cnn-based single image crowd counting and density estimation," *Pattern Recognition Letters*, vol. 107, pp. 3–16, 2018.

[11] L. Zhang, M. Shi, and Q. Chen, "Crowd counting via scale-adaptive convolutional neural network," in *WACV*, 2018, pp. 1113–1121.

[12] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, "Crowd counting via adversarial cross-scale consistency pursuit," in *CVPR*, 2018, pp. 5245–5254.

[13] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M. Cheng, and G. Zheng, "Crowd counting with deep negative correlation learning," in *CVPR*, 2018, pp. 5382–5390.

[14] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[16] M. A. Stricker and M. Orengo, "Similarity of color images," in *Storage and Retrieval for Image and Video Databases III*, 1995, vol. 2420, pp. 381–393.

[17] D. G Lowe, "Distinctive image features from scale-invariant keypoints," vol. 60, no. 2, pp. 91–110, 2004.

[18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, vol. 1, pp. 886–893.

[19] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *CVPR*, 2015, pp. 833–841.

[20] V. A. Sindagi and V. M. Patel, "Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *AVSS*, 2017, pp. 1–6.

[21] L. Wang, W. Shao, Y. Lu, H. Ye, J. Pu, and Y. Zheng, "Crowd counting with density adaption networks," *arXiv preprint arXiv:1806.10040*, 2018.

[22] S. Huang, X. Li, Z. Zhang, F. Wu, S. Gao, R. Ji, and J. Han, "Body structure aware deep crowd counting," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1049–1059, 2018.

[23] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "Decidenet: Counting varying density crowds through attention guided detection and density estimation," in *CVPR*, 2018, pp. 5197–5206.

[24] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS Workshop*, 2017.