A Dynamic-Attention on Crowd Region with Physical Optical Flow Features for Crowd Counting

1st Qian Wang
School of Computer Science
Shanghai Key Lab of Intelligent Information Processing
Fudan University
Shanghai, China
wangqian18@fudan.edu.cn

2nd Wenxi Li
The Academy for Engineering and Technology
Shanghai Key Lab of Intelligent Information Processing
Fudan University
Shanghai, China
wxli18@fudan.edu.cn

3rd Songjian Chen
The Academy for Engineering and Technology
Shanghai Key Lab of Intelligent Information Processing
Fudan University
Shanghai, China
sjchen18@fudan.edu.cn

4th Rui Feng
School of Computer Science
Shanghai Key Lab of Intelligent Information Processing
Fudan University
Shanghai, China
fengrui@fudan.edu.cn

***Abstract*—Crowd counting is widely used in various video surveillance applications. However, most of the existing approaches treat videos as a single frame, which increase redundant information and have low efficiency, due to ignoring the context history information of neighboring frames. In this paper, we propose a novel two-stream dynamic-attention network (DANet) to associate the temporal and spatial information. Specifically, the DANet includes two stages, one of which is to generate the region-attention map and the second is to refine the high-quality density map. In each stage, we develop a hierarchical fusion strategy to guide spatial attention, which can iteratively refine the region of crowds. Besides, the dynamic-attention module guided by the physical optical flow can be dynamically integrated into any network module to optimize the generation of features for improving the effect. Therefore, it can be plugged into many computer vision architectures. Finally, experimental results on three challenging benchmark datasets show that DANet outperforms most of the previous methods. Incorporating such dynamic-attention into a framework could boost the performance of end-to-end CNN-based methods.**

I. Introduction

Crowd counting is still a challenging problem especially the process of crowd video data. Most approaches for crowd counting process videos as the same way to single image, which treats videos to each frame independently with no consideration of the intrinsic temporal correlation among adjacent frames such as MCNN [1] proposed by Zhang et al. and CSRNet [2] introduced by Li et al. However, these CNN models are usually suitable for single image data which may increase lots of redundant information and have low efficiency.

Therefore, how to use contextual temporal information efficiently and accurately for crowd video data has become a worthwhile topic. In the past work, Miao et al. [3] proposed a spatial-temporal convolutional neural network, which incorporated spatial and temporal streams in an end-to-end manner. They used 2D and 3D Convolutional neural networks(C2D and C3D) to extract spatial-temporal features from a single frame and neighboring frames. Then a fusion network merged the features of C2D and C3D. Other researchers used the LSTM algorithm to extract video sequence information. For example, Xiong et al. [4] used a bidirectional convolutional LSTM model and utilize an approach in enforcing temporal consistency to use other information in the video sequence to achieve the crowd counting. Fang et al. [5] believed that LSTM-based methods may be affected by irrelevant history and interfere with prediction, so they used density maps directly to obtain historical information rather than used historical information which included identity information. However, the use of such temporal models often result in a large number of network parameters, and the structure of the model is too complex. These methods process fewer video frames per second and speed is slower.

Due to the constraints between the former and latter frames of the video data, these frames can provide information about the data itself. Therefore, our main task is to use the relation between the former and latter frames to optimize the model and improve the accuracy of recognition. Besides, we found the physical optical flow can roughly estimate the crowd region in the videos. An example of optical flow is shown in Fig. 1.

In this paper, inspired by the breakthroughs of the attention mechanism on other computer vision tasks, we propose a novel dynamic-attention network (DANet) to generate high-quality density maps for video data. Specifically, we use the movement information of the crowd in the video to generate the physical optical flow to obtain the approximate region information of crowds by calculating the size and direction of the optical

* Rui Feng is the corresponding author

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

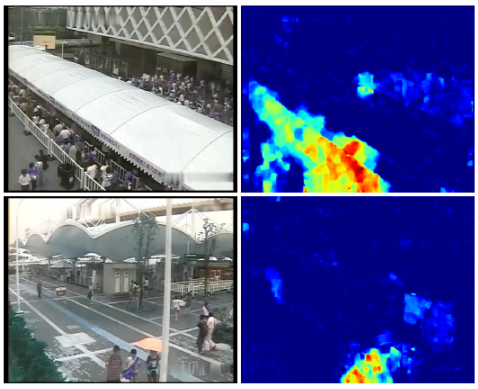

Fig. 1. The images on the left are the samples of the testing set in WorldExpo'10 dataset. On the right shows the optical flow generated by the sample.

flow, then converted into a gray image with a value from 0 to 1. The attention module of DANet is used to expand the attention weight of each pixel dynamically in the original image to generate the initial spatial attention map.

Compared with previous works, we use the optical flow map with physical features in the attention module, which makes the network can pay more attention to the region where the crowd is located. In addition, each training of the network can refine and iterate repeatedly, and fuse the useful historical information continuously between video frames to improve the accuracy of DANet. In general, our main contributions of this work are three-fold.

- We propose a novel two-stream framework DANet to improve the counting effect on video data. We use shallow optical flow features to assist generate attention maps. The DANet consists of two stages. In the first stage, DANet inputs the optical flow map and the initial video frame into two parallel networks respectively to extract features to obtain an initial spatial attention map. In the second stage, the iterative spatial attention map from the first stage and the original image are used as the input of the second stage two-streams network to generate the final density map.
- The DANet can be divided into spatial network stream and attention network stream. We propose a hierarchical fusion strategy to fuse different types of features. The features of spatial stream and attention stream are convolved between each layer to refine the region of interest gradually. The optical flow, which can capture the complementary motion information between consecutive frames, is used to generate the spatial attention map to help the network focus on learning the features of the crowd region.
- Extensive experiments and evaluations on multiple public benchmarks show that our proposed method achieves excellent performance. Specifically, our architecture achieves the best results and performance in the ShanghaiTech, UCF_QNRF datasets and almost all scenes in the WorldExpo'10 dataset.

II. Related Work

A. Regression-based methods for crowd counting

Lempitsky and Zisserman [6] proposed a method for regression of crowd density map. The density map not only counts the number of people but also generates the location information of the crowd. They used a fixed Gaussian kernel to normalize a dotted annotation, in order to generate a density map as ground truth. The count is the sum of each-pixel value in the ground truth. Moreover, MCNN [1] proposed a geometry-adaptive kernel to generate the ground truth, which determines the spread parameter σ based on the size of the head for each person within the image.

B. Deep learning for crowd counting

Recently, deep learning has greatly stimulated the progress in crowd counting. Inspired by the great success in deep learning on other tasks, Zhang et al. [7] focused on the CNN-based approaches to predict the number of crowds. In the same year, Wang et al. [8] proposed a deep convolutional neural networks regression model. However, they ignored the important perspective geometry of scene images and the fully connected layers in them throw away spatial coordinates. Zhang et al. [1] proposed a multi-column architecture (MCNN) where each column has a convolution kernel with different sizes for different scales. The multi-column architecture became a common way to deal with scale issues. For example, Switching-CNN [9] divided each image into non-overlapping patches and used a switch classifier to choose columns for patches. ACSCP [10] used non-overlapping patches and proposed adversarial loss instead of traditional Euclidean loss, due to Euclidean loss is sensitive to outliers and has the issue of image blur. Xiong et al. [4] proposed a model called ConvLSTM and it was the first time incorporating temporal-stream for crowd counting. More recently, Li et al. [2] proposed a model called CSRNet that used dilated convolution instead of last two pooling layers and outperformed most of the previous methods. They developed a single-column network to replace multi-column network, which has fewer parameters.

C. Attention for crowd counting

Previous research has shown that attention is important in human perception [11], [12]. Recently, some attention mechanisms of computer vision have been proposed [13], [14]. Attention mechanism is used to confer more weights on the features. DecideNet [15] is the first work to use channel attention in crowd counting. An attention module is used to adaptively decide the attention weight between two density maps for each pixel. However, training this model is difficult because of the multi-task learning. And DRSAN [16] developed a spatial transformer-based attention mechanism and refined the attended density map region with residual learning. The primary limitation of DRSAN is too many parameters.

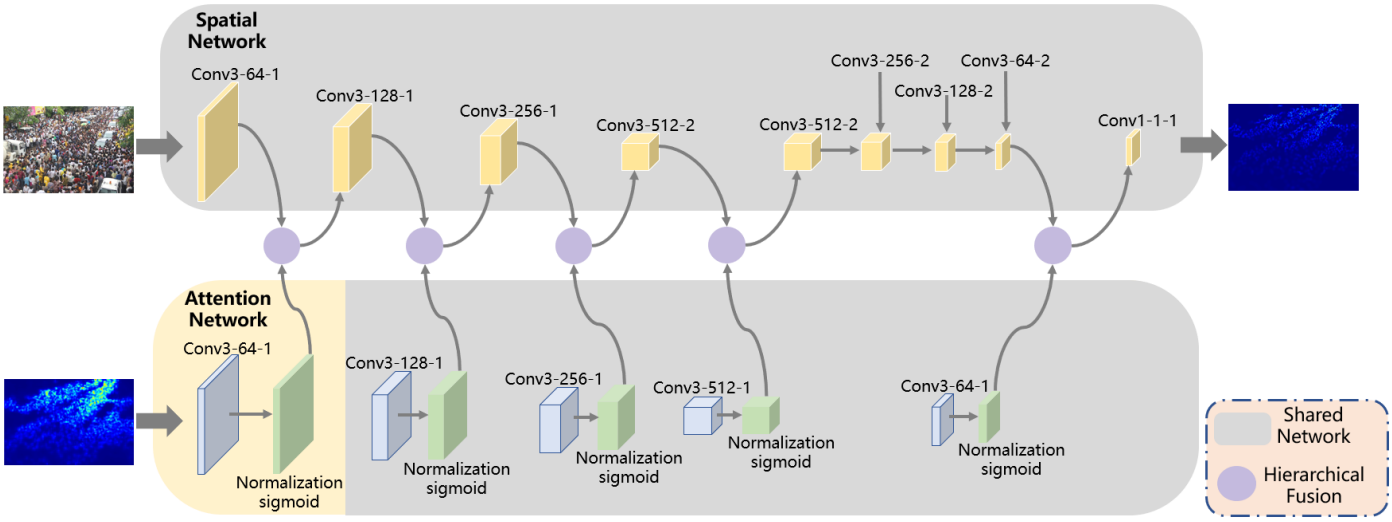

Fig. 2. Our convolutional attention architecture. DANet consists of spatial network(top row) and attention network(bottom row). The gray background area is the shared network, which used for two stages. The feature map of attention-stream is passed to normalization sigmoid and fusion with feature map of spatial-stream via an element-wise multiplication. In order to keep the output size and input size the same, all convolutional layers use padding. The parameters following conv are donated as kernel size, number of filters and dilation rate. The kernel size and stride of max-pooling are both 2.

D. Temporal stream for visual recognition

Inspired by the great success in video action recognition, the research on temporal-stream should also be considered in crowd counting. Recent works have shown that multi-frame dense optical flow can be used as a new feature to improve the accuracy of motion recognition. In order to extract more features of temporal-stream, Tran et al. [17] proposed a C3D network which operates on continuous RGB frames. However, this network needs a larger dataset and has more parameters, which limits the depth of the network. As we all know, LSTM is a unit that uses temporal feature and is widely used in natural language processing. In order to extend the LSTM to have convolutional structures, Shi et al. [18] proposed a model named ConvLSTM, which is already used for crowd counting [4]. Motivated by the FCAN [19], we use optical flow as guidance for spatial attention.

III. Proposed Method

A. Problem Statement

Suppose that we have a labeled training dataset of N images $\{(X_i, P_i)\}_{i=1}^{N}$, where P is a set of the pixels in an image X_i and each image is annotated with a set of point annotations $H_i = \{H_1, H_2, ..., H_{M_i}\}$, which denotes M_i 2D pixel locations of heads in the i-th image. Typically, let $D^{gt}(p) \geq 0, \forall p \in P$ be the i-th density map as ground truth for each pixel p in image, which is defined as,

$$D^{gt}(p) = \sum_{m=1}^{M} \mathcal{N}(p\ ;\ \mu_m, \delta^2) \tag{1}$$

$$= \sum_{m=1}^{M} \frac{1}{\sqrt{2\pi}\delta} \times e^{-(\frac{\|p-\mu\|_2^2}{2\delta^2})} \tag{2}$$

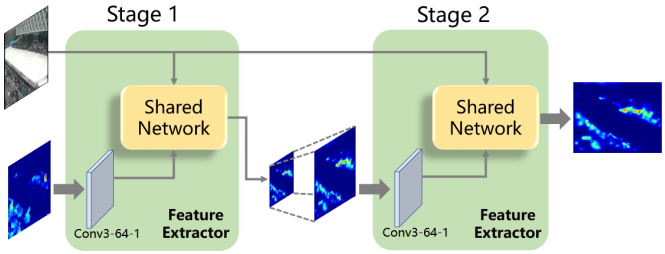

Fig. 3. Two-stage architecture. The attention network provides feature map to the spatial network, and feature map of attention-stream indicates the scores of different locations of the feature map of spatial network. Our approach is two-stage and the second stage utilizes the prediction from the first stage.

where $\mathcal{N}$ is a normalized Gaussian kernel. μ and δ^2 are mean and an isotropic covariance respectively.

The goal of our model is to learn a mapping function f (X): $X \in \mathbb{R}^{3\times W\times H} \mapsto Y \in \mathbb{R}^{W\times H}$ by minimizing the loss which calculates the difference between predicted density map D^{pre} of our model and the ground truth density map D^{gt} , in which f (X) means a deep network for image X, with W and H which are the width and height of the image and three channels.

$$Y = f(X) = W^* X \tag{3}$$

where W^* is a series of transformations. Crowd counting by density estimation methods regresses density maps by minimizing the pixel-wise Euclidean loss.

In this research, our purpose is to generate high-quality density maps for crowd video data with physical optical flow information. We propose a two-stream convolutional attention network to refine the learning crowd regional features, which consists of a space network and attention network. The architecture of DANet is illustrated in Fig. 2. The spatial network is CSRNet [2] based on VGG-16 [20] and the attention network

makes up of 3 × 3 convolutions, ReLU [21], and MaxPool in order. Hierarchical fusion strategy is an important part of the model, which is the fusion layer of spatial network and attention network.

B. Two-stage Crowd Counting

We propose an attention network to guide spatial attention, therefore the spatial attention map needs to be passed to the attention network. As shown in Fig. 3, we developed a two-stages architecture.

The first stage mainly includes the generation of attention map guided by physical optical flow map, which used to process other interference information in the video data. We calculate the magnitude and direction of the optical flow and use the HSV color model for visualization, where the direction corresponds to the hue value of the image and magnitude corresponds to the value plane. Then, we convert the HSV color model into a grayscale image with values in the range [0,1]. After that, we obtained the initial spatial attention map shown in Fig. 4(c).

Compared with continuous video data, single-frame images lack historical information. For this, our strategy is to initialize the optical flow map to a matrix with all numbers are one. This means that the spatial attention map at the initial stage has the same weights on all parts of the feature map of spatial-stream. After the first stage, the area of the crowd of interest is learned and the attention map is obtained.

Formally, Let I_att be the initial spatial attention map (optical flow or map where all values are 1) and x_att^1 be the feature map of the first layer in attention network, then

$$x\_att^1 = W_{stage1} \circledast I\_att \tag{4}$$

where $\circledast$ denotes convolution operation. W_{stage1} is the filter weights of the first layer in attention network for the first stage.

Shared network can use this feature map to generate a rough density map, which shows the attention of the head:

$$D_{initLR} = SharedNetwork(I\_rgb, x\_att^1) \tag{5}$$

where I_rgb is original image and D_{initLR} is initial density map with low resolution.

Because there are three pooling layers in the spatial network, the initial density map is one-eighth of the width and height of the original image. Therefore, in order to maintain the same size as the original image, we directly perform the interpolation method for upsampling:

$$D_{initHR} = U(D_{initLR}) \tag{6}$$

where shape of D_{initLR} is $H \times W \times C$ and shape of D_{initHR} is $rH \times rW \times rC$, r is the scale factor. D_{initHR} is initial density map with high resolution (spatial attention map). U is the upsampling function.

In the second stage, the iterative attention map obtained from the first stage and the original image are transmitted to the attention network and the spatial network respectively. This approach can enhance the crowd region, weaken other unrelated interference information and prevent the loss of important information at the shallow level. Attention maps at this stage are equivalent to the refined optical flow maps learned from the network.

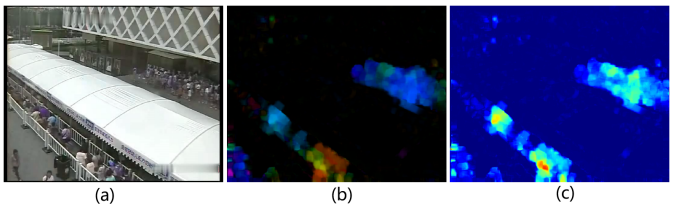

Fig. 4. (a) Visualization of the original image. (b) optical flow. (c) initial spatial attention map. The bright part of the initial spatial attention map represents the capture of more obvious motion, which means it could be pedestrians.

The features obtained from the spatial and attention network still processed by the proposed hierarchical fusion convolution method to complete the fusion learning features. Therefore, the feature map of attention-stream will guide spatial attention to improve the accuracy of prediction. In addition, the second stage is also to adapt to the static image which without optical flow information. The generation of the second stage is the final density map, which more accurate than initial density map (spatial attention map).

Similar to the first stage, the spatial attention map is passed to a convolution layer firstly:

$$x\_att^1 = W_{stage2} \circledast D_{initHR} \tag{7}$$

where W_{stage2} is the filter weights of the first layer in attention network for the second stage.

Afterwards, the feature map of the first layer in attention network and original image are passed to the shared network:

$$D_{final} = SharedNetwork(I\_rgb, x\_att^1) \tag{8}$$

where D_{final} is the final density map.

In addition, all layers of spatial and attention network are shared except for the first layer of attention network. The reason is that the input of the attention network is different on two stages, and the network needs different methods to extract features. In order to pass the feature map to the next layer which is shared, the number of filters between two layers is the same.

C. Hierarchical fusion strategy

DANet includes spatial and attention streams. The spatial network is the main network for feature extraction, and the attention network can assist the spatial network to adjust the weights of features in different regions. In addition, we introduce a novel hierarchical fusion strategy, which performing fusion convolution between spatial network and attention network. Hierarchical fusion strategy guides spatial attention via element-wise multiplication and utilizes the feature map of attention-stream effectively and refines the region of interest.

As is shown in Fig.2, DANet uses the hierarchical fusion strategy five times. We use the feature map of attention-stream to calculate the attention scores. The feature maps are

passed to the activation function and then multiplied with the corresponding regional features of spatial-stream. Therefore, crowd features will have high weights while other redundant features will be diluted.

Let x_rgb and x_att are the features learned by spatial-stream and attention-stream networks respectively. We propose a novel hierarchical fusion strategy that fuses the features of two streams for each layer.The hierarchical fusion includes regularization processing and learns the features of the crowd region by convolution layer processed by sigmoid function and element-wise multiplication.

Firstly, we normalize the feature map of attention-stream x_att by μ and σ which is the mean and variance of the feature map:

$$\hat{x}_{c,h,w} = \frac{x\_att_{c,h,w} - \mu}{\sigma} \tag{9}$$

where c is the channel of feature map, and w and h are the coordinates in feature map.

The normalization layer transforms the values of the feature map to a standard normal distribution, which $\mu = 0$ and $\sigma^2 = 1$. Then, we convert the values again by sigmoid function to get the attention score $a\epsilon\,[0, 1]$:

$$a_{c,h,w} = S(\hat{x}_{c,h,w}) \tag{10}$$

where S is the sigmoid function.

The attention score is used to correct the values of spatial-stream feature map:

$$x\_rgb_{att} = x\_rgb \odot a_{c,h,w} \tag{11}$$

where $\odot$ denotes element-wise multiplication.

IV. Experiments

In this section, we evaluate and compared DANet to the previous state-of-the-art methods in three datasets [1], [22], [7]. An ablation study is used to analyze the effect of different configuration and the flexible of attention network.

A. Datasets

ShanghaiTech [1] consists of two parts, including Part A and Part B. For Part A, the images contained in the dataset are from the internet. Part B is from busy streets of metropolitan areas in Shanghai. There are more congested scenes in Part A, and more sparse scenes in Part B. In Part A, 300 images are used for training and 182 images for testing. In Part B, 716 images are used for training and 400 images for testing.

WorldExpo'10 [7] contains 1,132 video sequences captured by 108 surveillance cameras during the Shanghai WorldExpo in 2010. The dataset is divided into training set and test set. There are 3380 annotated frames in training set, which are from 103 scenes. The test set consists of five scenes, which includes 120 frames respectively.

UCF-QNRF [22] is a new benchmark for crowd counting and the largest crowd dataset. The UCF-QNRF contains 1,535 annotated images with a total of 1.25 million people with centers of their heads annotated. The minimum and the maximum counts are 49 and 12865 respectively. Following, the training and test set consist of 1,201 and 334 images respectively.

Attention A	Attention B	Attention C
conv3-64-1-1	conv3-64-1-2	conv3-64-1-2
hierarchical fusion		
max-pooling		
conv3-128-1-1	conv3-128-1-2	conv3-128-1-2
hierarchical fusion		
max-pooling		
conv3-256-1-1	conv3-256-1-2	conv3-256-1-3
hierarchical fusion		
max-pooling		
conv3-512-1-1	conv3-512-1-2	conv3-512-1-3
hierarchical fusion		
4*conv3-64-2-1	conv3-512-2-0 conv3-256-2-1 conv3-128-2-1 conv3-64-2-1	conv3-512-2-3 conv3-256-2-1 conv3-128-2-1 conv3-64-2-1
hierarchical fusion		
conv1-1-1		

Fig. 5. Configuration of different attention network. In order to keep the output size and input size the same, all convolutional layers use padding. The parameters following conv are donated as kernel size, number of filters, dilation rate and number of layers. The kernel size and stride of max-pooling are both 2.

B. Evaluation Metric

The count error is measured by two metrics: Mean Absolute Error (MAE) and Mean Squared Error (MSE). The MAE is defined as

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|C_{x_i} - C_{x_i}^{gt}\right| \tag{12}$$

and the MSE is defined as

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|C_{x_i} - C_{x_i}^{gt}\right|^2} \tag{13}$$

where N is the number of images in test set, C_{x_i} is the estimated number of people, $C_{x_i}^{gt}$ is the number of crowds in ground truth.

C. Ablation Experiments

The Effect of Different Configuration: The depth of the attention network also needs to be considered. The features of the spatial attention map are fewer than the images captured by the camera, so the convolution layer should be less. We experiment with the attention network of three depths on ShanghaiTech dataset (Part A) and the configurations of them are shown in Fig. 5. Attention A has only 5 convolution layers. Attention C has the same architecture as spatial network, which has 16 convolution layers. Attention B has 11 convolution layers, which is between Attention A and Attention C. The evaluation results are shown in Table. I, and Attention A achieves the lowest error. Therefore, we use Attention A for the following experiments.

The Effect of Attention Network: In this subsection, we perform an ablation study to analyze the attention network on ShanghaiTech dataset (Part A) which is under highly congested

TABLE I
THE COMPARISON OF ARCHITECTURES ON SHANGHAITECH DATASET (PART A).

Architecture	*MAE*	*MSE*
Attention A	**64.3**	**96.1**
Attention B	66.1	102.7
Attention C	65.7	101.4

column1		column2		column3	
Spatial Network	Attention Network	Spatial Network	Attention Network	Spatial Network	Attention Network
conv9-16	conv9-16	conv7-20	conv7-20	conv5-24	conv5-24
hierarchical fusion		hierarchical fusion		hierarchical fusion	
max-pooling		max-pooling		max-pooling	
conv7-32	conv7-32	conv5-40	conv5-40	conv3-48	conv3-48
hierarchical fusion		hierarchical fusion		hierarchical fusion	
max-pooling		max-pooling		max-pooling	
conv7-16	conv7-16	conv5-20	conv5-20	conv3-24	conv3-24
hierarchical fusion		hierarchical fusion		hierarchical fusion	
conv7-8	conv7-8	conv5-10	conv5-10	conv3-12	conv3-12
hierarchical fusion		hierarchical fusion		hierarchical fusion	
merge layer					

Fig. 6. The configuration of MCNN with attention network.

scenes. We develop attention network for CSRNet and MCNN respectively. CSRNet with attention network is the DANet which is mentioned above. The configuration is shown in Fig. 2. The configuration of MCNN with attention network is shown in Fig. 6.

We intend to compare the results of original model and the model with attention network. Both of two comparison experiments are on ShanghaiTech dataset (Part A). As the experimental results shown in Table. II, attention network can effectively improve performance. In addition, attention network can improve performance on both CSRNet and MCNN. The reason for choosing these two networks for the ablation study is that the two networks are single-column architecture and multi-column architecture respectively. We evaluate the performance of our attention network on both architectures and show how hierarchical fusion strategy could be designed in two architectures. Therefore, the experimental results show the flexibility of our attention network, and the attention mechanism can be considered for expansion into more models.

D. Training Details

The spatial network is a VGG-based network, and the first 10 convolutional layers of spatial network are fine-tuned from a VGG-16 [20] which already pre-training. The initial values of the other filter weights are from a Gaussian initialization with 0.01 standard deviation. For the training scheme, we train DANet using the Pytorch library, optimizing to convergence using Stochastic gradient descent (SGD) with the fixed learning rate at 1e-7. For WorldExpo'10, we use two frames at interval 15 to calculate the optical flow. The Euclidean distance is chosen to calculate the difference between the generation and ground truth, which is similar to other works [2], [9]. The loss function is given as follow:

$$L(\Theta) = \frac{1}{2N}\sum_{i=1}^{N}\left\|Z(X_i, Y_i; \Theta) - Z_i^{gt}\right\|_2^2 \qquad (14)$$

where N is the batch size for training and $Z(X_i, Y_i; \Theta)$ is the generation of DANet with parameters shown as Θ. X_i is the sample of training set and Y_i is an initial spatial attention map. Z_i^{gt} is the ground truth of the input image X_i.

TABLE II
ABLATION EXPERIMENTS ON SHANGHAITECH DATASET (PART A). S IS DONATED AS SPATIAL NETWORK AND A IS ATTENTION NETWORK.

Basic	*Method*	*MAE*	*MSE*
2*MCNN [1]	S	110.2	173.2
	S+A	**102.7**	**156.4**
2*CSRNet [2]	S	68.2	115.0
	S+A	**64.3**	**96.1**

TABLE III
COMPARISON OF THE PERFORMANCE OF DIFFERENT MODEL ON SHANGHAITECH DATASET.

	Part A		**Part B**	
Method	***MAE***	***MSE***	***MAE***	***MSE***
MCNN [1]	110.2	173.2	26.4	41.3
Switching-CNN [9]	90.4	135.0	21.6	33.4
L2R [23]	72.0	106.6	13.7	21.4
ACSCP [24]	75.7	102.7	17.2	27.4
DRSAN [16]	69.3	96.4	11.1	18.2
IG-CNN [25]	68.5	116.2	10.7	16.0
CSRNet [2]	68.2	115.0	10.6	16.0
DANet(Ours)	**64.3**	**96.1**	**9.0**	**15.7**

E. Evaluations and Comparisons

We compared our DANet with numbers of state-of-the-art methods as listed in Table. III to V.

The results on the ShanghaiTech dataset are listed in Table. III. We use geometry-adaptive kernels to generate ground truth for both Part A and Part B. The proposed method is evaluated against seven recent approaches: MCNN [1], Switching-CNN [9], L2R [23], ACSCP [24], DRSAN [16], IG-CNN [25] and CSRNet [2], and we achieve the state-of-the-art performance. Compared with the result from CSRNet which DANet based on, DANet achieves 5.7% lower Mean Absolute Error (MAE) on Part A, and 15.1% lower MAE on Part B. The results indicate that our two-stream convolution attention can improve the performance.

The comparison results on the WorldExpo'10 datasets are presented in Table. IV. In training process, the frames are randomly cropped to the 1/2 size of the original image. Since the region of interest (ROI) map for each scene is provided, we are only focused on the heads under the ROI. For fair comparison, we use perspective maps to generate the ground truth maps which is similar to the work of Zhang et al. [7]. The relation is $\delta = 0.2M(x)$, where $M(x)$ denotes the number of pixels in the image representing one square meter at location x. The proposed method is evaluated against eleven recent approaches: Zhang et al. [7], MCNN, Switching-CNN,

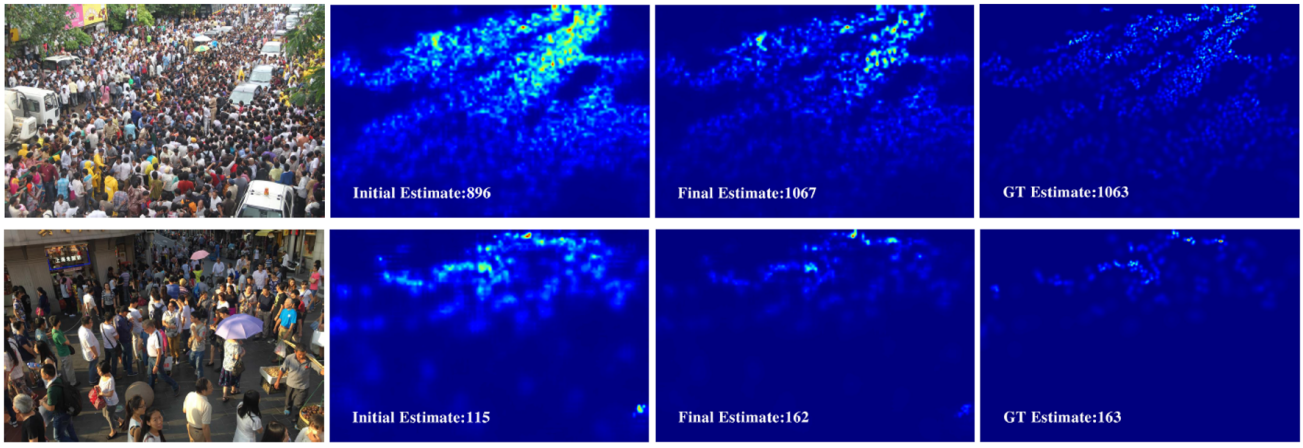

Fig. 7. First row: input images from ShanghaiTech dataset Part A and ground truth. Second row: input images from ShanghaiTech dataset Part B and ground truth. The initial estimate and the final estimate are the sum of initial density maps which are obtained in the first stage and the second stage respectively.

TABLE IV
COMPARISON OF THE PERFORMANCE OF DIFFERENT MODEL ON WORLDEXPO'10 DATASET.

Method	*Scene1*	*Scene2*	*Scene3*	*Scene4*	*Scene5*
Zhang et al. [7]	9.8	14.1	14.3	22.2	3.7
MCNN [1]	3.4	20.6	12.9	13.0	8.1
Switching-CNN [9]	4.4	15.7	10.0	11.0	5.9
ConvLSTM [4]	7.1	15.2	15.2	13.9	3.5
ACSCP [24]	2.8	14.05	9.6	**8.1**	2.9
D-ConvNet [26]	1.9	12.1	20.7	8.3	**2.6**
CSRNet [2]	2.9	**11.5**	8.6	16.6	3.4
ADMG [27]	3.8	14.5	11.7	17.9	3.5
SPANet [28]	3.4	14.9	15.1	12.8	4.5
RRSP [29]	2.2	11.1	11.3	15.8	2.8
PACNN [30]	2.3	12.5	9.1	11.2	3.8
DANet(Ours)	**1.8**	16.1	**7.7**	17.0	**2.6**

TABLE V
COMPARISON OF THE PERFORMANCE OF DIFFERENT MODEL ON UCF-QNRF DATASET.

Method	*MAE*	*MSE*
Idrees et al. [31]	315	508
MCNN [1]	277	426
Switching-CNN [9]	228	445
Idrees et al. [22]	132	191
DANet(Ours)	**115**	**186**

ConvLSTM [4], ACSCP, D-ConvNet [26], CSRNet, ADMG [27], SPANet [28], RRSP [29] and PACNN [30]. Specifically, our model gets the lowest MAE in three Scene which are Scene 1, Scene 3 and Scene 5.

On the UCF-QNRF dataset, we use geometry-adaptive kernels to generate ground truth for this dataset. The proposed method is evaluated against four recent approaches and experimental results are shown in Table. V.

F. Qualitative Analysis

In order to validate the two-stream convolutional attention mechanism of our method, we compared the generation of the two stages. As shown in Fig. 7, the visualization on density maps show that the generation of the first stage is the approximate location of the heads. Through the first stage, we can probably know the coverage of the crowd, but the scope is still relatively broad. This initial density map, which is as spatial attention map, is passed to the second stage, and a more refined location is achieved. Besides, the final density map is more similar to the ground truth. The improvement of experimental results compared with also indicate the effectiveness of our model.

When we tested on WorldExpo'10 dataset, we found that our model performed well in Scene 1, Scene 3 and Scene 5, but it was generally bad in Scene 2 and Scene 4. Even worse than the simple spatial network without using optical flow. In order to analyze, we visualize the optical flow of different scenes as shown in Fig. 1. We discover that there are more static crowd in Scene 2 and Scene 4, but the crowd in other scenes is always moving. In a static scene, optical flow will reduce the weight of these parts without movement. So the result is the same as what we saw via visualization, optical flow can extract features from videos which have dynamic crowd. While we also need to increase the robustness of our model for crowd with inapparent motion.

In addition, most datasets are only single images and they limit our work on crowd counting in video. WorldExpo'10 dataset has the surveillance video while the resolution is low.

A high-resolution video dataset can evaluate the performance of our model in videos better.

V. Conclusion

In this paper, a dynamic attention network (DANet) guided by physical optical flow features is proposed to associate the temporal and spatial information, which contains two stages. The network iteration in the second stage refines the attention feature map generated in the first stage, which represents the interested crowd region. Besides, we develop a hierarchical fusion strategy to combine each layer of region features from the attention network and the global features from the spatial network. Our approach shows state-of-the-art performances compared with other recent work. From experimental results, we can obtain several conclusions:

- Our attention module guided by the physical optical flow with temporal contextual information has positive effects on other network structures, which is available to obtain the interested crowd region.
- The dynamic-attention module can optimize the generation of features by integrated into other networks module dynamically.
- Our hierarchical fusion strategy flexibly combines the attention features and global features, which can effectively refines learning crowd features.

Acknowledgement

This work was supported by Military Key Research Foundation Project (No. AWS15J005) and National Natural Science Foundation of China (No. 61672165 and No. 61732004).

References

[1] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 589–597, 2016.

[2] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1091–1100, 2018.

[3] Y. Miao, J. Han, Y. Gao, and B. Zhang, "ST-CNN: Spatial-temporal convolutional neural network for crowd counting in videos," *Pattern Recognit. Lett*, pp. 113–118, july 2019.

[4] F. Xiong, X. Shi, and D. Yeung, "Spatiotemporal modeling for crowd counting in videos," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5151–5159, 2017.

[5] Y. Fang, B. Zhan, W. Cai, S. Gao, and B. Hu, "Locality-constrained spatial transformer network for video crowd counting," *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 814–819, 2019.

[6] V. Lempitsky and A. Zisserman, "Learning to count objects in images," pp. 1324–1332, 2010.

[7] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 833–841, 2015.

[8] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," pp. 1299–1302, 2015.

[9] D. Sam, S. Surya, and R. Babu, "Switching convolutional neural network for crowd counting," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4031–4039, 2017.

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

[11] M. Corbetta and G. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," vol. 3, no. 3, pp. 5245–5254, 2002.

[12] R. A. Rensink, "The dynamic representation of scenes," vol. 7, no. 1-3, pp. 17–42, 2000.

[13] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," pp. 3156–3164, 2017.

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

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

[16] L. Liu, H. Wang, G. Li, W. Ouyang, and L. Liang, "Crowd counting using deep recurrent spatial-aware network," *Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18*, 2018.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," pp. 4489–4497, 2015.

[18] S. Xingjian, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," pp. 802–810, 2015.

[19] A. Tran and L. Cheong, "Two-stream flow-guided convolutional attention networks for action recognition," pp. 3110–3119, 2017.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[21] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323, 2011.

[22] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 532–546, 2018.

[23] X. Liu, J. van de Weijer, and A. D. Bagdanov, "Leveraging unlabeled data for crowd counting by learning to rank," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7661–7669, 2018.

[24] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, "Crowd counting via adversarial cross-scale consistency pursuit," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5245–5254, 2018.

[25] D. Babu Sam, N. Sajjan, R. Venkatesh Babu, and M. Srinivasan, "Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3618–3626, 2018.

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

[27] J. Wan and A. Chan, "Adaptive density map generation for crowd counting," pp. 1130–1139, 2019.

[28] Z. Cheng, J. Li, Q. Dai, X. Wu, and A. G. Hauptmann, "Learning spatial awareness to improve crowd counting," pp. 6152–6161, 2019.

[29] J. Wan, W. Luo, B. Wu, A. Chan, and W. Liu, "Residual regression with semantic prior for crowd counting," pp. 4036–4045, 2019.

[30] M. Shi, Z. Yang, C. Xu, and Q. Chen, "Revisiting perspective information for efficient crowd counting," pp. 7279–7288, 2019.

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