An Indoor crowd detection network framework based on feature aggregation module and hybrid attention selection module

WenXiang Shen
Taiyuan City, Shanxi Province, China
SFlyswx@outlook.com

Pinle Qin
Taiyuan City, Shanxi Province, China
qpl@nuc.edu.cn

Jianchao Zeng
Taiyuan City, Shanxi Province, China
zjc@nuc.edu.cn

Abstract

In this paper, we present an indoor crowd detection network framework based on feature aggregation module and hybrid attention selection module (HSFA2Net). In order to better provide the details needed for small scale pupulation detection, we propose a novel feature aggregation module (FAM), which uses the idea of fusion and decomposition to aggregate contextual feature information. Since the indoor population feature and background feature overlap and the classification boundaries are not obvious, the proposed improved hybrid attention selection module (HASM) combines the selection mechanism with the previously proposed mixed attention module. Ultimately, we implement an indoor crowd detection network framework and achieve a recall rate of 0.92 and an F1 score of 0.92 on a public dataset SCUT-HEAD.

1. Introduction

The indoor crowd detection task, like the outdoor crowd couting task, has important research value in many real-world aspects such as teaching management, security alerting, event planning, *etc*. In recent years, there are two main deep leaning based approaches have been the mainstream of crowd couting, due to the powerful representation learning ability of convolutional neural networks (CNNs). One is the method of directly obtaining the count through regression [22, 30, 27], which can only predict the rough number of indoor crowd, but ignore other information such as behavior, movement trend and so on. In order to improve the above problems, researchers have recently inceased their focus on the use of detection [18, 9], which contains rich semantic information. Previously, human detection methods were mainly divided into two categories, one was pedestrian detection [34, 15] and the other was face detection [17, 10]. Nevertheless, face detection can only detect forward and side faces, and pedestrian detection has a poor ability to distinguish between occluded people. In this paper, we propose a strategy based on head detection to compenstate for the drawbacks of face and pedestrian detection, which contains both positioning and couting information. However, as illustrated in Figure 1, there are still two major challenges in head detection.

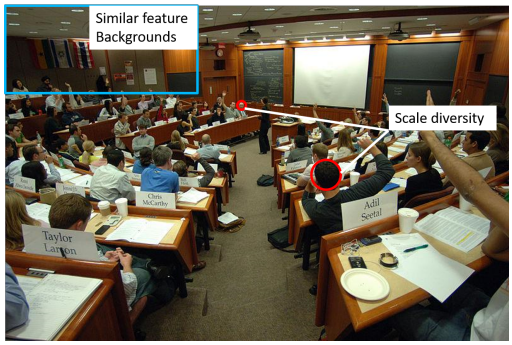

Figure 1. Indoor heads detection challenges.

The first major challenge is the diversity of head-scale dimension, especially for small-scale head. As illustrated in Figure 1, the scale of head can very significantly, which change from littleness to largeness. Such large scale pattern shifts usually bring grand challenges to head detection by a single CNN model, due to its fixed size of receptive fields. Remarkable progress has been achieved by learning a multiscale detector through designing multi-scale architectures [32] or aggregating multi-scale features [14, 21, 11]. Although these methods all generate a multi-scale detector with good performance, these methods combine the feature maps of multiple scales and directly use them to detect objects. The contribution rate of the merged features to objects

of different scales is not well considered. In this paper, we propose a simple yet effective method to mitigate the problem. The core idea is to decompose the aggregated features according to different scales of the head. Features containing rich context information are first fused through feature aggregation module to compensate for information detects between different layers of the CNNs. Then, according to the different scales of the head, the features including receptive field information are obtained by cascaded dilate convolution decomposition for final detection.

The second challenge is that in an indoor scene, the characteristics of the head are similar to the surrounding background features, especially in image patches that are far from the camera. As shown in Figure 1, the low-level features of the head's color, shape, and pixel mean overlap exactly with other objects in surrounding background. To overcome this problem, researchers have introduced attention mechanisms that force the network to focus on narrow areas associated with the target, ignoring interference from other background areas. Nevertheless, the previously proposed attention method only extracts the distribution of object information in the feature map from the channel dimension [7], the spatial dimension [28] or the mixed dimension [29], and does not consider the synergistic contribution and redundant information between the attention modules. In this paper, we propose a attention selection mechanism that works on a improved hybrid attention module (HASM). In particular, the proposed method consists of two steps. First, the input feature stream is respectively enhanced by the improved spatial and channel modules. Second, the enhanced feature is weighted and merged through the selection module.

Experiments are conducted on open indoor crowd detection datasets, including SCUT-HEAD [], Brainwash Dataset [24]. Extensive evaluations demonstrate superior performance over the prior arts. To summarize, the following are our main contributions:

- We propose an improved feature aggregation module that obtains accurate feature information by means of aggregate decomposition for objects of different scales, thereby improving detection performance.
- We propose a attention selection mechanism that acts on the improved hybrid attention module to preserve redundant attention while maintaining different levels of attention-enhancing performance.

2. Related Works

Due to the limited indoor space and the small number of people, unlike outdoor crowd counting, indoor crowd counting uses a detection-based approach that is more advantageous in terms of accuracy and subsequent application processing. Early methods for indoor crowd detection are mostly low-level feature extraction and detection, such as Haar [4], sobel [4] *etc*. Recently, with the development of deep learning, the mainstream crowd counting methods switch to CNN-based methods.

2.1. Deep ConvNet object detectors

Recently, there are two main ideas for object detection algorithms based on deep learning. The first type is a two-stage detector, such as R-CNN [5], Fast-RCNN [4], Faster-RCNN [20], R-FCN [2]. This type of method is to select the candidate region of the object and perform coarse screening, and then perform object classification and bounding box regression on the selected candidate regions. Although this type of method has high detection accuracy, it has a long inferred time and takes up high memory. The second category is single-stage detection algorithms including: SSD [14], YOLO [19]. This type of method produces the bounding box and category of the object directly from the image. The inference of such methods is fast, but the accuracy of object detection is especially poor for small objects.

2.2. Methods using multiple layers

Researchers have put plenty of efforts into improving the detection accruancy of objects with various scales no matter what kind of detector it is, either an single-stage detector or a two-stage one. To the best of our knowledge, there are mainly two strategies to tackle this scale-variation problem. The first one is featurizing image pyramids to produce semantically representative multi-scale features, such as: MTCNN [32], SNIP [23]. The second one is detecting object in the feature pyramid extracted from inherent layers within the network while merely taking a single-scale image including: MS-CNN [1], DSSD [3], FPN [11], RetinaNet [12].

2.3. Attention mechanism

Attention is a mechanism that mimics the processing tasks of the human brain. When the human brain processes tasks, it only focuses on the task itself and ignores other non-task interferences. The attention mechanism was first widely used in the field of natural language processing, such as natural language inference [16], text representation [25], sentence embedding [13] and so on. Meanwhile, the attention modules are also increasingly applied in the image vision field including: Image classification [8, 26], object detection [6], video tracking [35], super-resolution image generation [31] and so on. In these tasks, the attention mechanism is embedded in the different locations of the network as modules, and can be well migrated to other tasks.

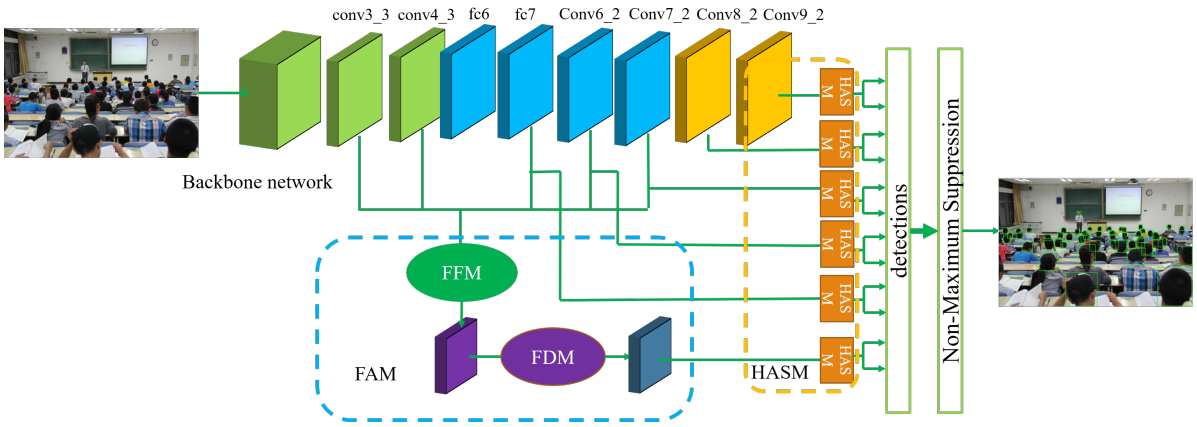

Figure 2. The overall architecture of HSFA2Net.

3. Proposed Method

3.1. Overview

The overall pipeline is depicted in Figure 2, consisting of two modules: 1) Feature aggregation module (FAM) presented in Sec. 3.2. We leverage multi-scale feature fusion to generate an initial feature map, which provides an accurate prediction and location on image. And then, according to the scale of different objects, the cascading hole convolution is used to construct the decomposition structure to obtain the feature information with different receptive fields; 2) Hybrid attention selection module (HASM) detailed in Sec. 3.3. The input feature information contains rich semantic information of the head and similar objects in the surrounding background. Therefore, it is necessary to first filter the target information through the improved mixed attention module proposed in this paper, and then use the proposed selection mechanism to reduce the redundant information. The whole network is end-to-end trainable and the training loss is presented in Sec. 3.4.

3.2. Feature Aggregation Module

Many previous multi-scale methods have proven to perform well in small object detection. Therefore, We follow the mainstream object detection methods by aggregating multi-scale feature. Empirically, we define an object as small when the area it occupies in images is smaller than 32×32 (the area is measured as the number of pixels in the segmentation mask). In the indoor crowd detection task, many people's head areas are smaller than this size. Taking VGG16 as the backbone network as an example, since the image is downsampled by convolution and pooling, the feature information of the small-sized object can only be transmitted to the middle layer Conv7 and disappears in the deeper layer afterwards. Therefore, instead of simply fusing multi-scale feature layers, we filter the feature layers that need to be fused based on the most remoteness that the smallest object feature information can propagate. Due to the lack of sufficient information, human beings are simultaneously assisting in the observation of small objects by means of surrounding features. We propose a feature decomposition method to mimic such human behavior for indoor crowd detection. Using the receptive field characteristics of the dilate convolution, we assign large receptive fields to small objects and small receptive fields to larger objects. Finally, we unify feature fusion and feature decomposition in a feature aggregation module.

The detailed structure of the feature fusion module is shown in the Figure 3. To reduce computational complexity and memory consumptionthe, feature fusion module first uses the 1×1 convolution to construct the bottleneck layer for channel normalization of the middle layer (conv4_3-conv7_2). In order to reduce the loss of spatial information, we directly reduce the size of the shallow conv3_3 using the bilinear down-sampling operation. Deconvolution can infer the activation information of the previous layer of convolution. Therefore, it can preserve the target semantic information well in the process of sampling on the feature layer, and reduce the interference of background semantic information. We use deconvolution to upsample the middle layer (fc7-conv7_2) features, then, element-level summation to fuse each of the upsampled middle-tier features to obtain a high-level semantic layer. Finally, we use the concate operation to connect the rich detail layer and the rich semantic layer to obtain the fusion features that contain global context information. Each of the convolution and deconvolution layers is followed by a ReLU layer.

In order to increase the information of small objects, we use the dilate convolution to construct the feature decomposition module, as shown in the Figure 4. Precisely, we use a 3×3 dilate convolution with a dilate ratio of 1, 2, 5 and a 1×1 standard convolution to decompose the input features

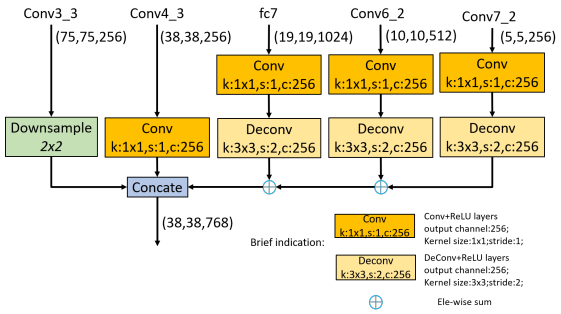

Figure 3. Feature Fusion Module architecture.

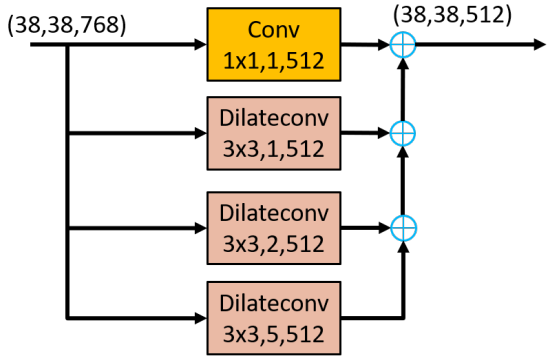

Figure 4. Feature Decomposition Module architecture.

and then use element-level summation to combine decomposition features for reconstructing feature layers containing different receptive field information.

3.3. Hybrid Attention Selection Module

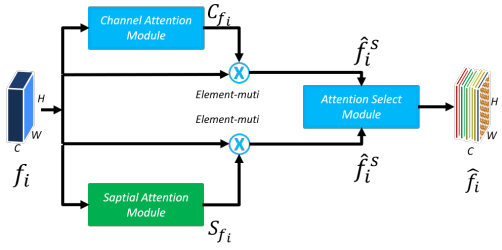

Figure 5. Hybrid Attention Selection Module.

We design a hybrid attention selection module for extracting key features. As shown in Figure 5, the input feature map $f_i \in \mathbb{R}^{H\times W\times C}$ extracts the channel attention map $C_{f_i} \in \mathbb{R}^{1\times 1\times C}$ with large target contribution rate through the channel attention module; Simultaneously, the spatial attention module is used to extract the two-dimensional spatial attention map $S_{f_i} \in \mathbb{R}^{H\times W\times C}$ to obtain the region with the highest correlation between the image and the target; Ultimately, we use the attention module in SEnet to

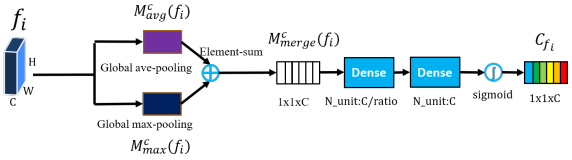

Figure 6. Channel Attention Module architecture.

build our attention selection module to obtain the final output $\hat{f}_i \in \mathbb{R}^{H\times W\times C}$. In the following, we will describe the improved channel attention module, spatial attention module and attention selected module in detail.

For different object, the feature maps of different channels have different contribution rates to key information. Channel attention is focused on the contribution of different channels. Therefore, this paper improves a structure for extracting the intrinsic relationship between channel and object, as shown in Figure 6. In order to learn only the contribution rate of different channels, the global average pooling method is generally used to compress spatial information. we also introduce the global maximum pooling operation at the same time. The global maximum pooling can obtain the most distinguishing features between channels, which can help to infer more detailed channel attention. First, channel attention module uses the global average pooling and the global maximum pooling to generate different spatial description features: $M^c_{ave} \in \mathbb{R}^{1\times 1\times C}$, $M^c_{max} \in \mathbb{R}^{1\times 1\times C}$. The merged channel description feature M^c_{merge} is then added by pixel-level addition. The merged channel description feature is fed into a multi-layer perceptron (MLP) to obtain the final channel attention map. In order to compress the parameters, this paper sets a compression ratio (dilate ratio), and through experiments, the parameter is finally set to 16. Finally, the process of attention extraction for the entire channel can be described as follows:

$$M^c_{merge}(f_i) = M^c_{ave}(f_i) + M^c_{max}(f_i) \quad (1)$$

$$C_{f_i} = \sigma(W_1 W_0(M^c_{merge}(x))) \quad (2)$$

In Equ. 1, $\sigma(.)$ is the sigmoid function. The principle of choice is that the channel attention extraction process belongs to a generalized two-class classification problem. The weight of the multi-layer perceptron: $W_0 \in \mathbb{R}^{C\times C/r}, W_1 \in \mathbb{R}^{C/r\times C}$, W_0 is activated with the nonlinear activation function ReLU.

The spatial attention is mainly to find the areas of the feature map that are important to the key information, which is a supplement to the attention of the channel. Since ordinary convolution operations are limited by the size of the convolution kernel, only the intrinsic association of features within the domain can be considered, and the correlation of similar features in the global region cannot be considered. Therefore, in order to obtain the contribution

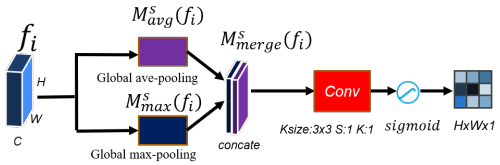

Figure 7. Spatial Attention Module architecture.

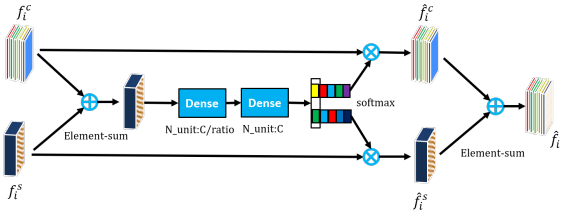

Figure 8. Attention Selected Module architecture.

of the global region to the key information, we design a improved spatial attention structure, as shown in Figure 9. First, the input feature map $f_i \in \mathbb{R}^{H\times W\times C}$ uses the global maximum pooling and the global average pooling to generate two new feature descriptions: $M^s_{ave} \in \mathbb{R}^{H\times W\times 1}$, $M^s_{max} \in \mathbb{R}^{H\times W\times 1}$, then fuses the new feature description through the concate operation, and finally obtains the spatial attention map through a standard convolution. The entire attention extraction process is described as follows:

$$M^s_{merge}(f_i) = [M^s_{ave}, M^s_{max}] \tag{3}$$

$$S_{f_i} = \sigma(f^{3\times 3} M^s_{merge}(f_i)) \tag{4}$$

Through the spatial attention module and the channel attention module, the input features will separately obtain different attention-enhancing information. Enhanced features will introduce redundant information if they are directly fused. Therefore, inspired by the gating idea in SEnet, we design the attention selection module. As shown in Figure 8, we merge the enhanced features through the concate operation, and then, the sequeeze and excitation operation is used to obtain weights for different attention modules. Finally, output features are obtained through pixel-level summation. As shown in the Figure 9, the left side is the original image, the right side is the shallow feature map of the SSD and the partial shallow layer feature of the HSFA2Net with the attention mechanism added. It can be seen that the attention module designed in this paper enhances the semantic information and detail location information of the target area in the feature map.

3.4. Training Loss

The whole network is end-to-end trainable, which involves two loss functions: 1) location loss L_{loc} is used to calculate the difference between the object position predicted by the network and the ground truth label; 2) classification loss L_{cls} is used to indicate the degree of matching between the predicted object category and the ground truth label. The final loss function L for the whole network is the combination of the above two losses given by

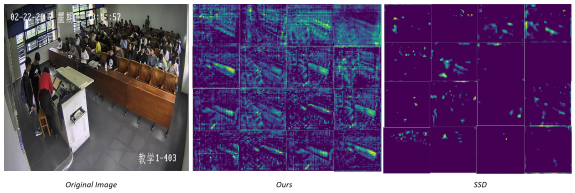

Figure 9. Our method, SSD part shallow feature map visualization.

$$L = \frac{1}{N} L_{cls} + \alpha L_{loc} \tag{5}$$

In Equ. 5, the hyperparameter α is a balance factor used to balance the effects of classification loss and location loss on the final structure. Here we select $\alpha = 1$ based on multiple experiments. $N = 0$ is the default number of frames matched. If $N = 0$, the set loss is 0.

The location loss in this paper is also the same as the previous detection methods using the smooth L1 loss. The deteail is shown below:

$$\begin{aligned} L_{loc}(x,l,g) &= \sum_{i\in Pos}^{N} \sum_{m\in\{cx,cy,w,h\}} x^k_{ij} smooth_{L1}(l^m_i - \hat{g}^m_j) \\ where \quad \hat{g}^{cx}_j &= (g^{cx}_j - d^{cx}_i)/d^{\omega}_i \quad \hat{g}^{cy}_j = (g^{cy}_j - d^{cy}_i)/d^{h}_i \\ \hat{g}^{cx}_j &= (g^{cx}_j - d^{cx}_i)/d^{\omega}_i \quad \hat{g}^{cy}_j = (g^{cy}_j - d^{cy}_i)/d^{h}_i \end{aligned} \tag{6}$$

where (cx, cy) is the center point of the detection box, and ω and h define the detection box's width and height. We use l to represent the predicted box position value and g to represent the position value of the real box.

We adopt multi-class confidence loss as the classification loss function of this paper.

$$\begin{aligned} L_{conf}(x,c) &= \sum_{i\in Pos}^{N} x^p_{ij} \log(\hat{c}^p_j) - \sum_{i\in Neg} \log(\hat{c}^0_i) \\ where \quad \hat{c}^p_i &= \frac{\exp(c^p_i)}{\sum_p \exp(c^p_i)} \end{aligned} \tag{7}$$

4. Experiments

4.1. Experimental Setup

All models are trained on Tesla M40 GPU. Before the training phase, we use random horizontal flip, random

brightness and data normalization as data preprocessing. Our method uses the SSD300 [33] pre-trained parameters on MSCOCO for parameter initialization. In the training phase, we use the stochastic gradient descent optimizer, the momentum is set to 0.9, and the weighting regularization parameter is set to 0.0005. The initial learning rate is set to 0.001. When training 80k, the learning rate drops to 1e-4, and after training for 20k, the learning rate was finally adjusted to 1e-5.

4.2. Datasets and Evaluation Metrics

SCUT-HEAD. This is a large-scale head detection dataset, which follows the standard of Pascal VOC, including 4405 images labeld with 111251 heads. This dataset consists of two parts. PartA includes 2000 images sampled from monitor videos of classrooms in an university with 67321 heads annotated. PartB includes 2405 images crowd from Internet with 43930 heads annotated. Both PartA and PartB are divided into training and testing parts.

Brainwash. This dataset contains 91146 heads annotated in 11917 images. We use this dataset only for testing.

Evaluation metrics. We employ three standard metrics, *i.e.*, Recall (R), Precision (P), and F1 score (F).

4.3. Experimental Comparisons

The proposed method outperforms all the other competing methods on all the benchmarks. The quantitative comparison with the state-of-the-art methods on these two datasets.

SCUT-HEAD. Table 1 compares our method with best performing methods on the SCUT-HEAD. Compared with other algorithms, we have a high improvement under various evaluation indicators, and each performance index is higher than 0.9. Table 2 shows the performance comparison of our method and other methods on small head detection. In the field of indoor crowd detection, our method reaches the SOTA level.

Brainwash. We also compare our method on Brainwash dataset in Table 3. Our method also achieves state-of the-art performance on this dataset compared with several baselines including context-aware CNNs local model (Con-local) [**?**] , SSD, R-FCN, and FRN [18].

4.4. Ablation study

Middle feature layers Choice. In order to verify the rationality of the feature layer selection used for fusion, we design different feature layer combination ablation experiments. The dataset uses the SCUT-HEAD PartA section, and all experimental conditions are the same as before. As shown in the Table 4, it can be found that the shallow layer conv3_3 to the middle layer conv7_2 is used for fusion, and the final performance index is the best, which proves the rationality of the structure design of the feature fusion module.

Method	PartA			PartB		
	P	R	F1	P	R	F1
Faster-RCNN[20]	0.86	0.78	0.82	0.87	0.81	0.84
YOLOv3[19]	0.91	0.89	0.89	0.74	0.67	0.70
SSD[14]	0.87	0.68	0.76	0.80	0.66	0.72
R-FCN(ResNet-50)	0.87	0.78	0.82	0.90	0.82	0.86
R-FCN+FRN[18]	0.89	0.83	0.86	0.92	0.84	0.88
HSFA2Net(proposed)	**0.93**	**0.92**	**0.92**	**0.95**	**0.91**	**0.93**

Table 1. Comparison of other methods and our method on SCUT-HEAD Dataset

Average scale	0∼10px			10∼20px		
Method	P	R	F1	P	R	F1
SSD	0.08	0.06	0.07	0.48	0.65	0.48
R-FCN	0.12	0.10	0.11	0.53	0.76	0.62
R-FCN+FRN	0.17	0.19	0.18	0.83	0.76	0.79
Ours	**0.23**	**0.22**	**0.22**	**0.87**	**0.84**	**0.86**

Table 2. Comparison of other methods and our method on small head detection

Method	Con-local	SSD	R-FCN	FRN	Ours
AP	44.5	80.2	84.8	88.1	**89.2**

Table 3. Comparison of other methods and our method on Brainwash Dataset

Method	fusion layer	P	R	F1
HSFA2Net	conv3_3-conv7_2	**0.92**	**0.90**	**0.92**
HSFA2Net	conv4_3-conv7_2	0.92	0.89	0.90
HSFA2Net	conv4_3-conv6_2	0.90	0.88	0.89
HSFA2Net	conv3_3-conv6_2	0.88	0.85	0.86

Table 4. Feature Fusion Module Structure Ablation Experiment Result

Attention module selection. In order to verify the rationality of our proposed dual attention structure design, five different structures are designed: the first is the baseline of the no-attention module; the second is the channel attention module introduced in the SENet; the third is the channel attention module introduced in this paper; the fourth is the spatial attention module introduced in this paper; the last one is the dual attention module containing channel attention and space attention. According to the results shown in the Table 5, the hybrid attention selection module designed in this paper can better improve the performance of the network.

4.5. Qualitative Results

We show some visualization results of our method and other methods, as shown in Figure 10, Compared with the

Method	P	R	F1
HSFA2Net(baseline)	0.88	0.86	0.87
HSFA2Net+SEBlock	0.94	0.88	0.89
HSFA2Net+CAM	0.92	0.87	0.89
HSFA2Net+SAM	0.93	0.88	0.90
HSFA2Net+CAM+SAM	0.92	0.90	0.92
HSFA2Net+CAM+SAM+ASM	**0.93**	**0.92**	**0.92**

Table 5. Fusion Attention Module Structure Comparison Experiment Results

visualization results of other methods, it can be seen that the HSFA2Net proposed in this paper solves the problem of multi-scale and the detection problem with similar object and environmental characteristics.

5. Conclusion

In this paper, we present an indoor crowd detection network framework based on feature aggregation module and hybrid attention selection module. The feature aggregation module can aggregate the context information and apply the context information to the object detection according to the scale of the head. The proposed hybrid attention selection module is used to enable the network to learn to distinguish between targets and surrounding similar object features and to reduce redundant information through a selection mechanism. Extensive experiments on two popular datasets demonstrate that the proposed method achieves consistent and significant improvements over the previous methods. HSFA2Net also shows the noteworthy generalization ability to untraining datasets, demonstrating the effectiveness of HSFA2Net in real applications.

References

[1] Zhaowei Cai, Quanfu Fan, Rogério Schmidt Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *ECCV*, 2016.

[2] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, 2016.

[3] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C. Berg. Dssd : Deconvolutional single shot detector. *CoRR*, abs/1701.06659, 2017.

[4] R. Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, Dec 2015.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, June 2014.

[6] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei. Relation networks for object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, June 2018.

[7] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, June 2018.

[8] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*, 2015.

[9] C. Le, H. Ma, X. Wang, and X. Li. Key parts context and scene geometry in human head detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1897–1901, Oct 2018.

[10] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Ji-Lin Li, and Feiyue Huang. DSFD: dual shot face detector. *CoRR*, abs/1810.10220, 2018.

[11] T. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, July 2017.

[12] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, Oct 2017.

[13] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017.

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pages 21–37, 2016.

[15] J. Mao, T. Xiao, Y. Jiang, and Z. Cao. What can help pedestrian detection? In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6034–6043, July 2017.

[16] Fandong Meng. Neural machine translation by jointly learning to align and translate. 2014.

[17] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis. Ssh: Single stage headless face detector. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4885–4894, Oct 2017.

[18] D. Peng, Z. Sun, Z. Chen, Z. Cai, L. Xie, and L. Jin. Detecting heads using feature refine net and cascaded multi-scale architecture. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2528–2533, Aug 2018.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, June 2016.

[20] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2017.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

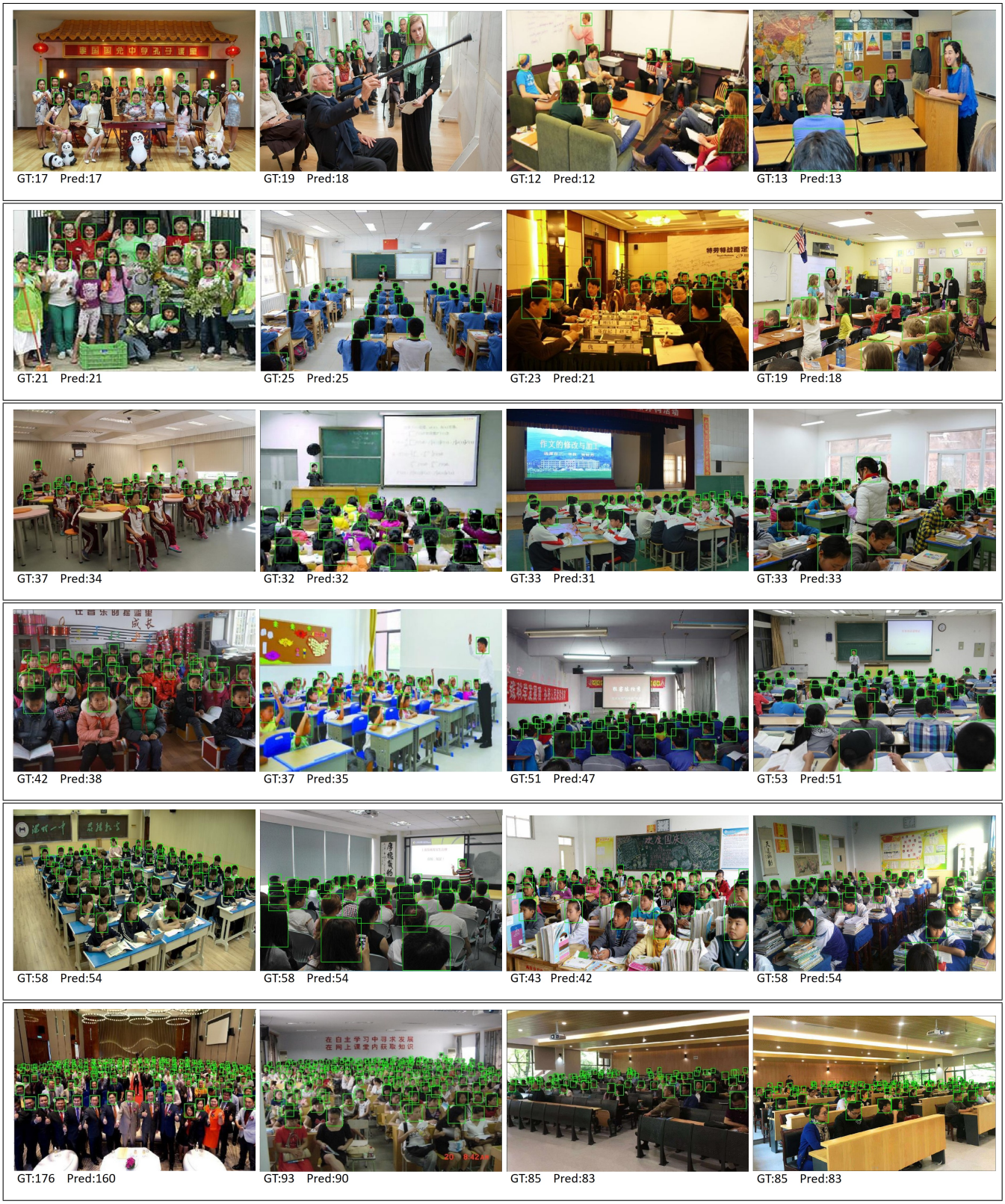

Figure 10. Qualitative results of HSFA2Net on the validation set of the SECUT-HEAD dataset.

[22] V. A. Sindagi and V. M. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1879–1888, Oct 2017.

[23] B. Singh and L. S. Davis. An analysis of scale invariance in object detection - snip. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3578–3587, June 2018.

[24] Russell Stewart, Mykhaylo Andriluka, and Andrew Y. Ng. End-to-end people detection in crowded scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2325–2333, 2016.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[26] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6458, July 2017.

[27] J. Wang, L. Wang, and F. Yang. Counting crowd with fully convolutional networks. In *2017 2nd International Conference on Multimedia and Image Processing (ICMIP)*, pages 210–214, March 2017.

[28] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, June 2018.

[29] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, pages 3–19, 2018.

[30] F. Xiong, X. Shi, and D. Yeung. Spatiotemporal modeling for crowd counting in videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5161–5169, Oct 2017.

[31] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *CoRR*, abs/1805.08318, 2018.

[32] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.

[33] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Single-shot refinement neural network for object detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4203–4212, 2018.

[34] S. Zhang, J. Yang, and B. Schiele. Occluded pedestrian detection through guided attention in cnns. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7003, June 2018.

[35] Z. Zhu, W. Wu, W. Zou, and J. Yan. End-to-end flow correlation tracking with spatial-temporal attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 548–557, June 2018.