DAReFet:Detecting Indoor Heads using Dual Attention And Refactoring Feature

Pinle Qin[1], Wenxiang Shen[1], and Jianchao Zeng[1⋆]

No.3 Xueyuan Road, Jiancao District, Taiyuan City, Shanxi Province, China
qpl@nuc.edu.cn sflyswx@outlook.com zjc@nuc.edu.cn

Abstract. In this paper, we propose a novel single-stage indoor head detection network using feature reconstruction and dual attention (DAReFet) that can be used for indoor population counting. To improve the performance of small head detection, we propose a novel Feature Reconstruction Module (FRM) for enhancing the semantic information of small object in shallow feature maps, which firstly combines the shallow and middle layer feature maps with the fusion strategy, and then uses multi-scale dilate convolution to decompose the fusion feature map for forming a reconstructed shallow feature map. Due to the characteristics of the heads similar to the background, we propose a novel Dual Attention Module (DAM), which greatly enhances the classification performance of the network on the target and background by fusing spatial and channel attention. Our method has achieved a recall rate of 0.90 and an F1 score of 0.92 on SCUT-HEAD, which achieves the state-of-the-art performance in the field of indoor crowd detection. DAReFet also leads to a runtime of 40 ms/image on a GPU.

Keywords: head detection · feature reconstruction · multi-scale dilate convolution · dual attention.

1 Introduction

Indoor population counting, as a typical application in the field of computer vision, has important value in many aspects such as statistical analysis of human traffic business data, teaching management and public safety. There are two main ideas for indoor crowd counting. One is the regression [26, 36], and the other is based on the detection method [22, 11]. Regression-based methods can only predict population density and get a rough result. The detection-based method can simultaneously obtain accurate locating information and population statistics. Therefore, the detection-based method has great interest among researchers. The most direct strategy for detecting crowd is face detection [19, 12] and pedestrian detection[41, 17]. However, these two methods have poor performance in indoor crowd detection. Face detection can only detect forward and side face, which limits that the camera's power. Due to the complexity of the

⋆ correspond author

indoor scene, many body parts overlap each other, pedestrian detection does not solve this problem well. Head detection does not have these limitations and can be well adapted for indoor crowd locating and counting. However, there are still two major challenges in head detection in indoor scenes, as shown in Figure 1.

The first challenge is the diversity of head-scale dimensions, especially for small-scale heads. In order to solve this problem, two strategies have been proposed by researchers using images of different scales to improve the problem. For example, as illustrated in Figure 2(a), Zhang et al. [39] proposed a construct called MTCNN, which obtains input images of different scales by direct downsampling and feeds them into the trained detection network for prediction, finally outputs the object position through non-maximum value suppression (NMS) [21]. Recently, Singh et al. [27] propose to create image pyramid of different sizes in training and testing, running a detection network on each graph, while retaining only those outputs whose size is within the specified range. Finally, the object location and category are also obtained through NMS operations. Obviously, these solutions will greatly increase memory and computational complexity. Therefore, the method based on the feature pyramid aims to solve the problem more effectively. These methods are to detect object in a feature pyramid extracted from the input image, which can be exploited at both training and testing phases. As illustrated in Figure 2(b), Liu et al. proposed a single-stage detection network (SSD) [16] using multi-scale features for the first time, which directly and independently uses two layers of the backbone (i.e. VGG16 [25]) and four extra layers obtained by stride 2 convolution to construct the feature pyramid; and then He et al. [13] proposed a detection network based on feature pyramid structure, which constructs the feature pyramid by fusing the deep and shallow layers in a top-down manner, as show in Figure 2(c). Compared to the previous strategy, the second strategy utilizes less memory and is less time consuming, and can also be embedded as a component in different object detection networks. Therefore, in this paper, we propose a new feature reconstruction structure (FRM) based on the idea of feature reuse, as illustrated in Figure 2(d). Due to continuous downsampling, the small object area has been reduced to 1×1 pixel size in the middle layer feature map.Therefore, FRM utilizes upsampling to combine only the shallow and intermediate layer features to form a new fusion feature map, and then uses the multi-scale dilate convolution pyramid to perform feature decomposition to form new small object detection branches, and assists a multi-level detection with the other object detection branches generated by the original feature layer.

The second challenge is that the head features are easily similar to the background features. Therefore, another key challenge is how to detect target object from complex backgrounds. Recent research has explored the attention mechanism used to guide the detection network to focus only on the target. Hu et al. [9] propose a channel attention mechanism that automatically acquires the importance of each feature channel through learning, and then enhances useful channel information. Wang et al.[35] propose a Non-local layer that can well capture the dependence between pixels at distant locations, which is equivalent

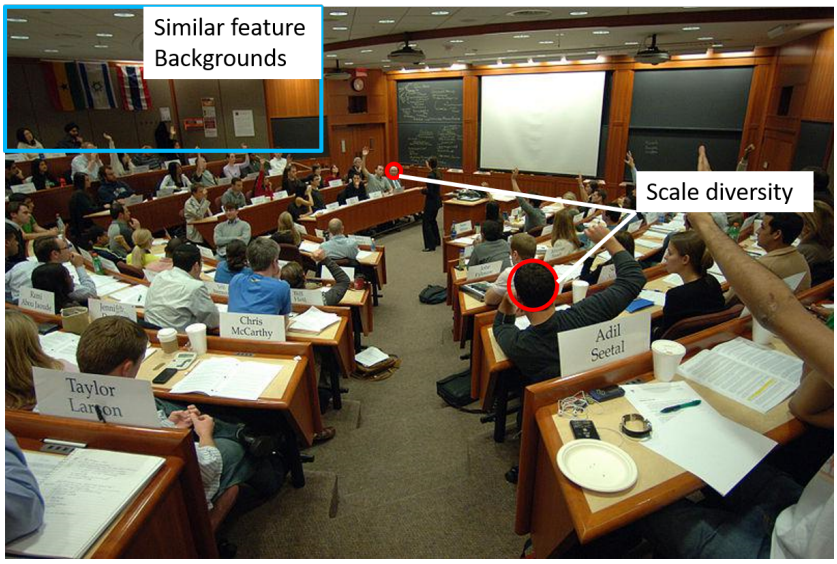

Fig. 1. Indoor heads detection challenges.

to paying attention to image position. But these methods only focus on a single type of attention.Different from the previous methods, we propose a dual attention mechanism to obtain channel attention and spatial attention separately, and then combine them for better detection.

To summarize, the following are our main contributions:

- We propose a novel Feature Reconstruction Module (FRM) that uses the upsampling and dialte convolution pyramid to fuse and decompose the intermediate layer and shallow features to obtain a new feature map containing rich small object information.
- We propose a novel Dual Attention Module (DAM) that includes spatial positional attention and channel attention to enhance the network's ability to detect object.
- DAReFet uses VGG16 lightweight feature extraction network as the backbone, combined with our proposed FRM and DAM to form a single-stage multi-level detection end-to-end network, making full use of the contribution of different layers of the network to different scale object, achieving 25fps Real-time inference speed on a GPU.

2 Proposed Method

2.1 General Architecture

DAReFet proposes to solve the problem of multi-scale head detection and complex background interference. The overall structure of our method is shown in

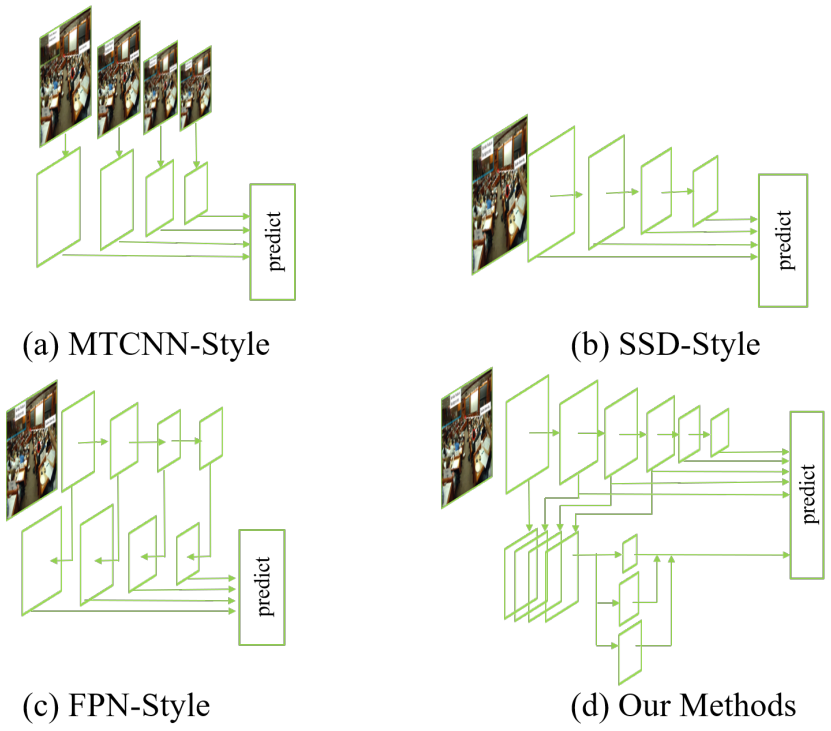

Fig. 2. Different Multi-Scale Detection Structures.

Figure 3. It uses VGG16 as the backbone, which embeds the Feature Reconstruction Module (FRM) proposed in this paper and Dual Attention Module (DAM), to extract features containing contextual information from the input image, and then similar to SSD[16], produces dense bounding boxes and category scores based on the learned features, followed by the non-maximum suppression (NMS) [20] operation to produce the final results. Inspired by the idea of fusion and decomposition, we have designed a Feature Reconstruction Module (FRM) that includes a Feature Fusion Module (FFM) and a Feature Decomposition Module (FDM) to solve the problem of multi-scale object detection, especially small object detection. Firstly, the shallow and intermediate layer feature fusion is realized by nearest neighbor downsampling and deconvolution to obtain rich context information. Then, the dilate convolution is used to simulate the multi-scale receptive field to decompose the fusion feature without changing the feature layer size. We introduce a Dual Attention Module (DAM) in each detection branch to solve the background feature interference problem. DAM consists of a cascading approach between channel attention and spatial positional attention. In the following we will discuss the design of the feature reconstruction module and the dual attention module in detail.

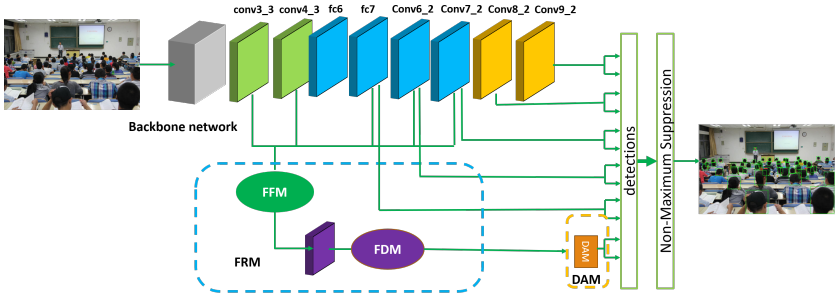

Fig. 3. Our method Architecture.

2.2 Feature reconstruction module (FRM)

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). As shown in Figure 1, due to the position of the camera, the small person's head is dominant in the indoor scene image. if the area of a head is about 30×30, we could obtain the fine details only on the shallow layers within the convnet (conv3_3-conv7_2). The representation of fine details for the head will bocome weaker and weaker on the following several layers and will be totally lost on the coarse,semantic deepest layer. We intend to make full use of the shallow layers with rich fine details and the relatively deep layers with semantic information as well as some fine details for small object. Therefore, we first designed the feature fusion submodule (FFM) to fuse the deep layers of rich semantic information with the shallow layers of rich detail information, as shown in Figure 4. Then, we combine the dilate convolution of different dialte ratios to design a feature decomposition sub-module that simulates the multi-scale receptive field, which is used to obtain a wider range of information for small object detection, as illustrated in Figure 5. Finally, we use cascaded methods to transform the features of different layers in the original network through fusion and decomposition.

Feature fusion submodule (FFM) As we can see in Figure 4, to reduce computational complexity and memory consumptionthe, FFM 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 downsampling 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

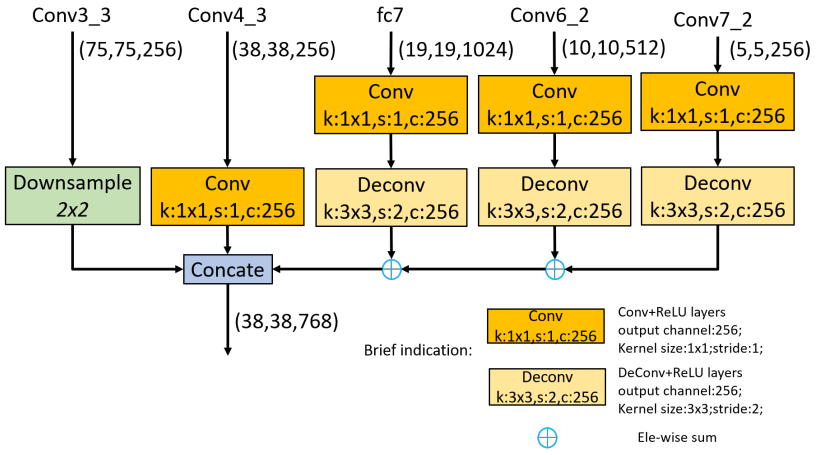

Fig. 4. Feature Fusion Module architecture.

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.

Feature decomposition submodule (FDM) A fusion feature layer containing rich context information, which is output by the FFM, will be used to generate a new small object detection branch. We introduce the dialte convolution into our proposed feature decomposition structure (FDM) to obtain more object information for small object detection. As shown in Figure 5, we use a 3×3 dilate convolution with a dilate ratio of 1, 2, 5 to decompose the input features. In addition, to detect smaller objects, we add a 1×1 standard convolution. Finally, we use element-level summation to combine decomposition features for reconstructing feature layers containing different receptive field information.

2.3 Dual Attention Module (DAM)

If the network can only use the feature map with high contribution rate to key object and abandon the feature map with low contribution rate to key object, it will greatly improve the positioning and recognition of the object. So, this paper designs a hybrid attention module for extracting key features. The overall structure is shown in Figure 6, and the input feature map $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ extracts the channel attention map $\mathbf{F}(\mathbf{x}) \in \mathbb{R}^{1 \times 1 \times C}$ with large target contribution rate through the channel attention module. By means of cascading, the spatial attention module is used to extract the two-dimensional spatial attention map $\mathbf{G}(\mathbf{x}) \in \mathbb{R}^{H \times W \times C}$ to obtain the final output. The representation of the entire

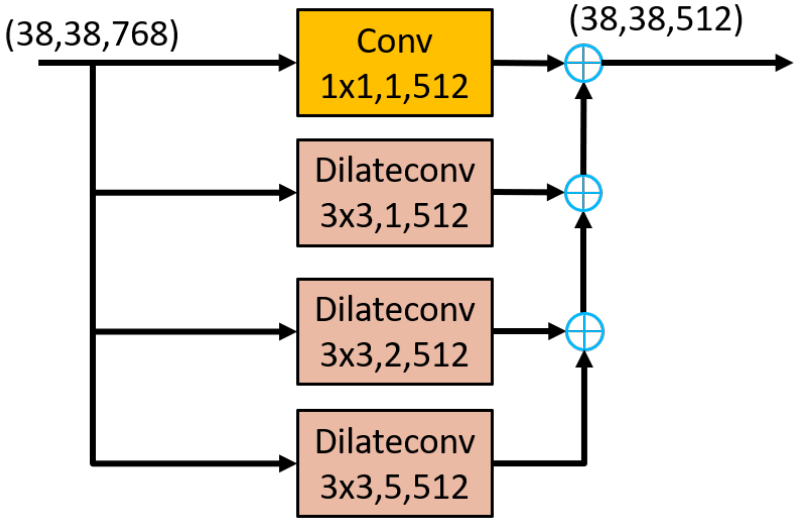

Fig. 5. Feature Decomposition Module architecture.

attention extraction process is as follows:

$$\mathbf{Z}(\mathbf{x}) = \mathbf{G}(\mathbf{F}(\mathbf{x}) \otimes \mathbf{x}) \otimes \mathbf{F}(\mathbf{x}) \tag{1}$$

In the Equ 1, $\otimes$ is the element multiplication. During the element multiplication process, the attention map is broadcast to the feature maps of different channels and different regions. The final output $\mathbf{Z(x)}$ contains both spatial attention and channel attention. As shown in the Figure 7, 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 DAReFet 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. In the following, we will describe the channel attention mechanism and spatial attention mechanism module in detail.

Channel Attention Module (CAM) Matthew et al. [37] propose that each channel feature map can be regarded as a feature detector. 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 to critical information. Therefore, this paper proposes a structure for extracting the intrinsic relationship between channel and object, as shown in Figure 8. In order to learn only the contribution rate of different channels, the global average pooling method is generally used to compress spatial information. For example, Bolei et.al [42] propose to use global average pooling to obtain target detection

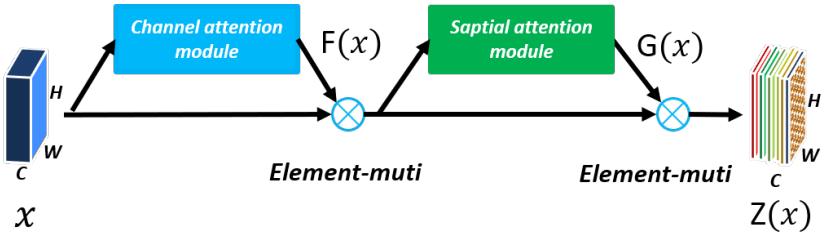

Fig. 6. Dual Attention Module architecture.

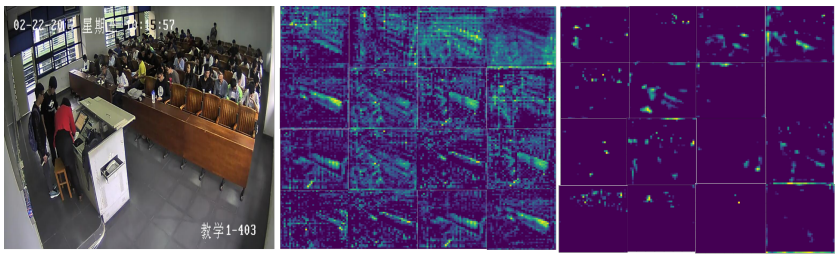

Fig. 7. Our method, SSD part shallow feature map visualization.

candidate regions; SENet uses global average pooling to obtain contribution rate information for the feature map channel. Different from their strategy, this paper introduces 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, use the global average pooling and the global maximum pooling to generate different spatial description features: $\mathbf{M^c_{ave}} \in \mathbb{R}^{1\times1\times C}$, $\mathbf{M^c_{max}} \in \mathbb{R}^{1\times1\times C}$. The merged channel description feature $\mathbf{M^c_{merge}}$ 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:

$$\mathbf{M^c_{merge}}(\mathbf{x}) = \mathbf{M^c_{ave}}(\mathbf{x}) + \mathbf{M^c_{max}}(\mathbf{x}) \tag{2}$$

$$\mathbf{F}(\mathbf{x}) = \sigma(\mathbf{W1}(\mathbf{W0}(\mathbf{M^c_{merge}}(\mathbf{x})))) \tag{3}$$

In Equ 3, $\sigma(.)$ 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: $\mathbf{W0} \in \mathbb{R}^{C\times C/r}$, $\mathbf{W1} \in \mathbb{R}^{C/r\times C}$,$\mathbf{W0}$ is activated with the nonlinear activation function ReLU.

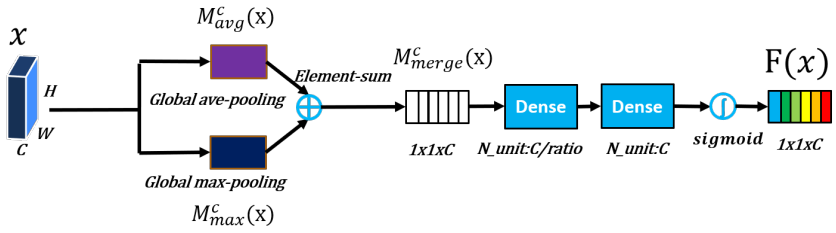

Fig. 8. Channel Attention Module architecture.

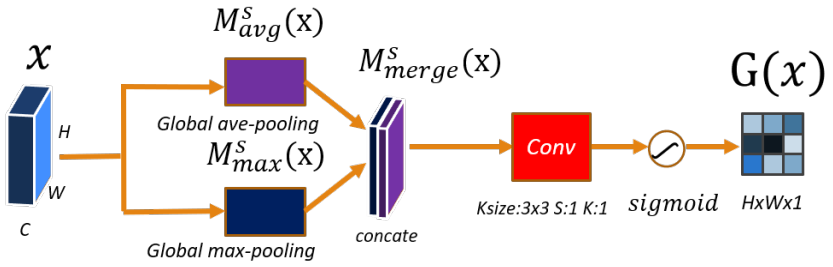

Fig. 9. Spatial Attention Module architecture.

Spatial Attention Module (SAM) 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 of the global region to the key information, inspired by the non-local network [35]. We design a novel spatial attention structure, as shown in Figure 7. First, the input feature map $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ uses the global maximum pooling and the global average pooling to generate two new feature descriptions: $\mathbf{M^s_{ave}} \in \mathbb{R}^{H\times W\times 1}$,$\mathbf{M^s_{max}} \in \mathbb{R}^{H\times W\times 1}$, 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:

$$\mathbf{M^s_{merge}}(\mathbf{x}) = [\mathbf{M^s_{ave}}, \mathbf{M^s_{max}}] \tag{4}$$

$$\mathbf{G}(\mathbf{x}) = \sigma(\mathbf{f^{3\times 3}}\mathbf{M^c_{merge}}(\mathbf{x})) \tag{5}$$

In Equ 5, $\sigma(.)$ is the sigmoid function, and $\mathbf{f^{3\times 3}}$ represents the standard convolution operation of 3×3.

2.4 Loss Function

Object detection includes both classification and regression tasks, so we need to build a multitasking loss function. The loss function in this paper is defined as the weighted sum of location loss and classification loss, as follows:

$$\mathbf{L}(\mathbf{x}, \mathbf{c}, \mathbf{l}, \mathbf{g}) = \frac{\mathbf{1}}{\mathbf{N}}(\mathbf{L_{conf}}(\mathbf{x}, \mathbf{c}) + \alpha \mathbf{L_{loc}}(\mathbf{x}, \mathbf{l}, \mathbf{g})) \tag{6}$$

In Equ 6 , 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. This article uses the center point coordinates of the box(cx, cy) and the width(ω) and height(h) parameters to define the position of a target frame in the image. Since smoothL1 is smoother than the L2 regression loss, the loss of smoothL1 between the predicted box(l) and the real label(g) is used as the location loss. As shown below:

$$\begin{aligned} L_{loc}(x, l, g) &= \sum_{i \in Pos}^{N} \sum_{m \in \{cx, cy, w, h\}} x_{ij}^{k} smooth_{L1}(l_i^m - \hat{g}_j^m) \\ where \quad \hat{g}_j^{cx} &= (g_j^{cx} - d_i^{cx})/d_i^{\omega} \quad \hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^{h} \\ \hat{g}_j^{cx} &= (g_j^{cx} - d_i^{cx})/d_i^{\omega} \quad \hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^{h} \end{aligned} \tag{7}$$

The classification loss uses the softmax multi-class loss, as shown in the following equation:

$$\begin{aligned} L_{conf}(x, c) &= \sum_{i \in Pos}^{N} x_{ij}^{p} \log(\hat{c}_j^p) - \sum_{i \in Neg} \log(\hat{c}_i^0) \\ where \quad \hat{c}_i^p &= \frac{\exp(c_i^p)}{\sum_p \exp(c_i^p)} \end{aligned} \tag{8}$$

3 Experiments

3.1 Experimental Setup

All models are trained on Tesla M40 GPU. Before the training phase, we used random horizontal flip, random brightness and data normalization as data preprocessing. Our method uses the SSD300 [40] pre-trained parameters on MSCOCO [30] 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.

3.2 Datasets

SCUT-HEAD [22]: 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 [28]: This dataset contains 91146 heads annotated in 11917 images. We use this dataset only for testing.

3.3 SCUT-HEAD Dataset Result

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. Table2 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.

Method	PartA			PartB		
	P	R	F1	P	R	F1
Faster-RCNN[24]	0.86	0.78	0.82	0.87	0.81	0.84
YOLOv3[23]	0.91	0.89	0.89	0.74	0.67	0.70
SSD[16]	0.87	0.68	0.76	0.80	0.66	0.72
R-FCN(ResNet-50)[4]	0.87	0.78	0.82	0.90	0.82	0.86
R-FCN+FRN[22]	0.89	0.83	0.86	0.92	0.84	0.88
DAReFet(proposed)	**0.92**	**0.90**	**0.92**	**0.94**	**0.91**	**0.92**

Table 1. Comparison of other methods and our methods 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	**0.22**	**0.21**	**0.21**	**0.86**	**0.83**	**0.85**

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

3.4 Brainwash Dataset Result

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) [31], SSD, R-FCN [4], and FRN[22].

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 methods on Brainwash Dataset

3.5 Ablation study: Middle feature layers Choice

In order to verify the rationality of the feature layer selection we used for fusion, we designed 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	fusion layer	P	R	F1
DAReFet	conv3_3-conv7_2	**0.92**	**0.90**	**0.92**
DAReFet	conv4_3-conv7_2	0.92	0.89	0.90
DAReFet	conv4_3-conv6_2	0.90	0.88	0.89
DAReFet	conv3_3-conv6_2	0.88	0.85	0.86

Table 4. Feature Fusion Module Structure Ablation Experiment Result

3.6 Ablation study: Attention design

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 mixed attention mechanism designed in this paper can better improve the performance of the network.

method	P	R	F1
DAReFet(baseline)	0.88	0.86	0.87
DAReFet+SEBlock	0.91	0.88	0.89
DAReFet+CAM	0.92	0.87	0.89
DAReFet+SAM	0.93	0.88	0.90
DAReFet+CAM+SAM	**0.92**	**0.90**	**0.92**

Table 5. Fusion Attention Module Structure Comparison Experiment Results

3.7 Qualitative Results

We show some visualization results of our method and other methods, as shown in Figure 10, Compared with the visualization results of other methods, it can be seen that the DAReFet proposed in this paper solves the problem of multi-scale and the detection problem with similar object and environmental characteristics.

4 Conclusion

In this paper, we propose a novel single-stage indoor heads detection networks called DAReFet, which is mainly used to detect indoor populations, and obtain the final population count statistics based on the test results. DAReFet consists of two modules: Feature Reconstruction Module (FRM) and Dual Attention module (DAM). FRM is used to solve the problem of multi-scale object, while DAM is used to solve the problem of similar object and background features. Finally, it is unified into a single-stage object detection framework to achieve end-to-end training and prediction. The inference speed on the GPU reaches 25fps. Our method achieves an F1 score of 0.92 and a recall rate of 0.90 on a standard data set. The method in this paper is flexible and simple, and can also be migrated to other object detection tasks.

References

1. Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *ECCV*, 2016.
2. Z. Cai, Q. Liu, S. Wang, and B. Yang. Joint head pose estimation with multi-task cascaded convolutional networks for face alignment. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 495–500, Aug 2018.
3. J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, 2016.
4. J. Dai, Y. Li, K. He, and J. Sun. R-FCN: object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 379–387, 2016.
5. C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. Dssd : Deconvolutional single shot detector. *CoRR*, abs/1701.06659, 2017.

Fig. 10. Qualitative results of DAReFet on the validation set of the SECUT-HEAD dataset.

6. R. Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, Dec 2015.
7. 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.
8. 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.
9. 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.
10. M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *NIPS*, 2015.
11. 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.
12. J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang. DSFD: dual shot face detector. *CoRR*, abs/1810.10220, 2018.
13. 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.
14. 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.
15. Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017.
16. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. 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.
17. 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.
18. F. Meng. Neural machine translation by jointly learning to align and translate. 2014.
19. 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.
20. A. Neubeck and L. J. V. Gool. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR 2006), 20-24 August 2006, Hong Kong, China*, pages 850–855, 2006.
21. A. Neubeck and L. Van Gool. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 850–855, Aug 2006.
22. 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.
23. 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.

24. 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.
25. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.
26. 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.
27. 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.
28. R. Stewart, M. Andriluka, and A. 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.
29. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017.
30. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):652–663, 2017.
31. T. Vu, A. Osokin, and I. Laptev. Context-aware cnns for person head detection. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2893–2901, 2015.
32. 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.
33. 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.
34. P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. W. Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*, pages 1451–1460, 2018.
35. 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.
36. 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.
37. M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 818–833, 2014.
38. H. Zhang, I. J. Goodfellow, D. N. Metaxas, and A. Odena. Self-attention generative adversarial networks. *CoRR*, abs/1805.08318, 2018.
39. 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.
40. S. Zhang, L. Wen, X. Bian, Z. Lei, and S. 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.

41. 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.
42. B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.
43. 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.