DSCA-Net:Indoor Head Detection Network Using Dual stream information and channel attention

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

College of Big data,The North University of China
No.3 Xueyuan Road,Jiancao District,Taiyuan City,Shanxi Province,China

Abstract. In this paper, we propose a novel indoor head detection network using dual stream information and mutil-attention that can be used for indoor population counting. In order to solve the problem of target scale diversity in indoor human head detection, especially the problem of small-scale human head, we propose a dual stream information flow structure to enrich the positioning and category semantic information of small-scale objects. At the same time, we propose a kind of The structure of the channel-attention mechanism is used to enhance the ability of the network to identify small-scale objects. Our method has achieved a recall rate of 0. 91 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.

Keywords: pedestrian detection · object detection · attention mechanism · deep learning.

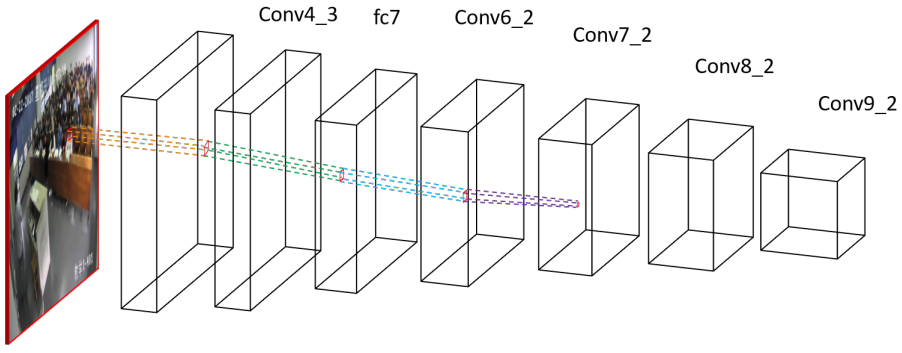

Fig. 1: Small objects feature spread in each layer.

⋆ Corresponding author

1 Introduction

The indoor crowd count is the same as the outdoor crowd count, and has important research value in public safety, business analysis and other application fields. In the outdoor crowd counting task, we often use the regression method[**?**] to solve the problem. However, this method can only estimate the number of people based on the premise of sacrificing positioning accuracy. The crowd count in the indoor scene, due to the low concentration and relatively fixed background environment, therefore, we can use the detection method to more accurately count the number of people. As previously suggested by researchers, face detection[7, 3] and pedestrian detection[21, 6] are directly used to solve this task. However, due to the installation angle of the indoor surveillance camera, people's physical characteristics and facial features are incomplete in the crowd, which is easy to be misidentified and detected by the network. Ultimately, we plan to use the idea of head detection for indoor crowd detection and counting, and this also avoids privacy issues caused by face detection. There are also two challenges in head detection.

The first challenge is the multi-scale problem of the human head. Due to the installation position of the indoor surveillance camera, the scale of objects far from the camera drops dramatically. Therefore, how to enhance the robustness of the network to scale is currently of concern to scholars. Previous research has achieved great success in this area, from the earliest direct use of multi-scale data to the current popular feature pyramid approach. For example, Zhang et al. [19] proposed a network 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 suppression (NMS) [9]; Singh et al. [14] 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; Liu et al. [5] proposed a single-stage detection network (SSD) [5] using multi-scale features and four extra layers obtained by stride 2 convolution to construct the feature pyramid; He et al. [4] 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. All of these methods suggest that shallow layers have rich details such as location, while deep layers have rich class semantic information. Therefore, the combination of shallow and deep features will improve the detection of small-scale objects on the network. However, this does not consider that objects of different scales require different levels of semantic information. Just like human vision, when we observe different objects, the objects entering the receptive field and the surrounding information are different. Therefore, this paper proposes a multi-level structure for constructing fusion feature information of different receptive fields for objects of different scales.

The second challenge is that the head features are easily coincident with the background features surrounding the room. Moreover, since we construct multi-level fusion information, this will inevitably introduce unnecessary noise, which makes it difficult to use shallow detectors with weak classification ability.

Recently, some scholars have explored the attention mechanism to enhance the attention of the network to the target, thus enhancing the discriminating ability of the network. Hu et al. [2] 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. [16] propose a Non-local layer that can well capture the dependence between pixels at distant locations, which is equivalent to paying attention to image position. Therefore, based on our task, we have designed a multi-attention mechanism module to enhance the ability of the detection branch to classify the object.

To summarize, the following are our main contributions:

- We propose a novel dual stream information module (DSM) that uses the upsampling and dialte convolution pyramid to fuse and decompose the intermediate layer and shallow features to enhance robustness to multiple scales.
- We propose a channel-attention module (CAM) that includes channel attention to enhance the network's ability to detect object.
- DSCA-Net uses VGG16 lightweight feature extraction network as the backbone, combined with our proposed DSM and MAM 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.

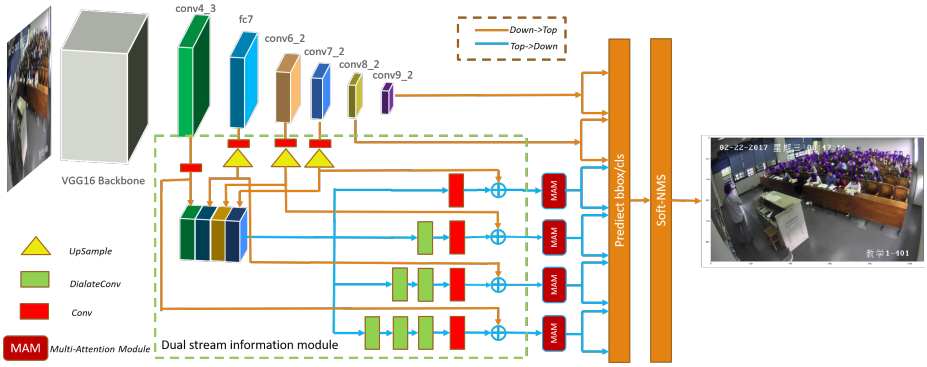

Fig. 2: Our method Architecture.

2 Proposed Method

2.1 General Architecture

The overall structure of DSCA-Net is shown in Figure 2. It uses VGG16 as the backbone, which embeds the multi-level imformation structure (MLIS) and multi-attention module (MAM), to extract features containing contextual information from the input image, and then similar to SSD, produces dense bounding

boxes and category scores based on the learned features, followed by the non-maximum suppression (NMS) [8] operation to produce the final results. Our multi-level information flow structure contains top-down and down-top two-way information. And according to the objects of different scales, we use the receptive field characteristics of the dilate convolution to generate different object detection features. The multi-focus module we proposed consists of a SE block. Its purpose is mainly to enhance the discriminating ability of the regenerated detection branch, and to reduce redundant information generated by the multi-stream fusion structure. In the following we will discuss the design of DSM and CAM in detail.

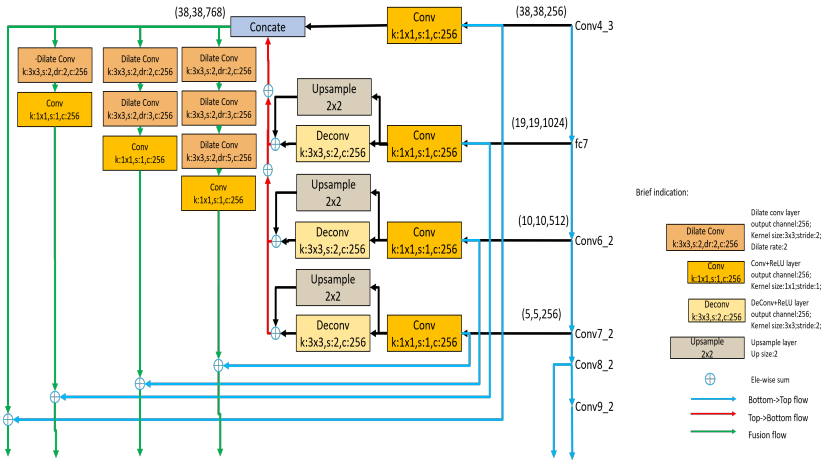

Fig. 3: Dual Stream information Module.

2.2 Dual Stream information Module (DSM)

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). Due to the position of the camera, the small person's head is dominant in the indoor scene image. As shown in Figure 1, 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 on the following several layers and will be totally lost on the 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, when we design the top-down information flow, we only combine the information in the middle layer and the shallow layer. After obtaining the top-down fusion information, we are not simply outputting

the detection results directly. Instead, we use the dilate convolution to construct the detection branches of different receptive fields. We expect the network to use different receptive fields to output objects of different scales. Finally, we use the summation method to combine the bottom-up information flow to construct global feature attention in the different receptive branches. This will enable the shallow and deep layers of the network to have rich position and discriminant semantic information on the basis of the original. In this way, our network can adaptively output results from different branches for objects of different scales.

Bottom to Top Flow (BTF) As indicated by the blue line on the far right of Figure 3, we first use bottom-top information to detect large-scale objects, which is consistent with our improved SSD network. We believe that large-scale objects only need to pay attention to the features that can reflect the properties of the objects when they are detected, and do not need the assistance of the surrounding background. Therefore, we only use the bottom-up feature information here. In order to enhance the category semantic information of small objects, we also use bottom-up information for subsequent information fusion to provide the global feature attention that the network originally learned, as shown in the middle blue line in Figure 3. In addition, we are using ordinary convolution operations to achieve the above operations.

Top to Bottom Flow (TBF) As the previous research shows, when detecting small objects, the shallow layer of the network lacks the necessary class semantic information, and the deep layers of the network lose the necessary object location information due to continuous downsampling. Therefore, in order to compensate for the shortcomings of the network when detecting small objects, we constructed a top-down path, as shown by the red line Figure 3. In order for the top and bottom information to blend, we need an upsampling operation. The direct upsampling of the image may result in the loss of detail of the object, while the deconvolution operation may infer the possible activation map of the previous layer, but lacks the reasoning of global information. Therefore, we connect the two operations through the sum operation to form complementary feature information for top-down propagation.

dilate pyramid decomposition (DPD) When we get top-down information, we do not directly integrate with the bottom-up information, but first perform multi-scale dilate pyramid decomposition. We are inspired by human vision. When human beings observe different objects, the range of receptive fields required is different, so we only need to pay attention to the characteristics of the objects themselves when we pay attention to large objects. When we pay attention to small objects, if we assist the surrounding related background, it will be greatly Improve detection rate. If we use the pooling operation directly to increase the receptive field, we will lose the detail features as before. Therefore, we use the receptive field characteristics of the dilate convolution to decompose the

receptive field into four different scales for better detection of objects of different scales. According to the method proposed by Wang et al., we also use stacked hole convolution to eliminate the grid effect.

2.3 Channel-Attention Selected Module (CAM)

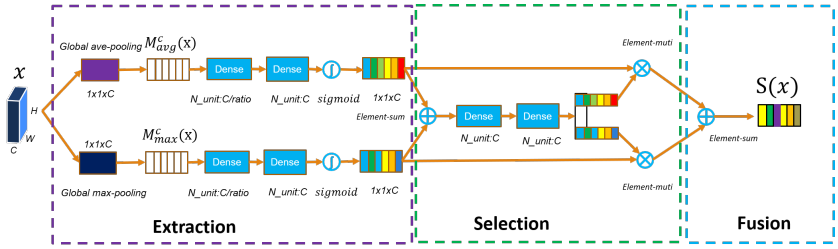

Fig. 4: Channel-Attention Selected Module.

The visual attention mechanism is a unique mechanism for signal processing in the human brain. Human vision selects some local key areas of interest by observing the global image, and then pays more attention to these areas to obtain more detailed information and suppress other Useless information. Generally, for the feature map of $C \times H \times W$, the $H \times W$ plane weights of the spatial attention are different, and the C weights are the same; the C weights of the channel attention are different, and the weights of the $H \times W$ planes are the same. Channel attention pays attention to "what is", and spatial attention focuses on "where." We only focus on the Bbox category generated based on the fused feature information, so here we design a channel attention module with a selection mechanism. As shown in Figure 4, we divide the design attention module into three parts, extract features, select channels, and merge channels.

In the feature extraction part, we use the global average pooling and global max pooling methods to compress the image into the c dimension. As we all know, the main sources of feature extraction error are: 1. The estimated value variance of the neighborhood size is limited (receptive field is small); 2. The error of the estimated mean value caused by the convolutional layer parameter error. In general, average-pooling can reduce the first type of error, retain more background information of the image, max-pooling can reduce the second error and retain more texture information.

In the selection phase, we re-multiplexed the form of SE Block and used the gating mechanism to achieve the choice of global average pooling and global max pooling. Since both the previous global average pooling and global max pooling generate different channel weights, and simple additions, redundant weight information is generated, thereby causing unnecessary judgment of the network. Therefore, we have designed this selection structure to maximize the retention of key detail features and global semantic information of objects.

Finally, based on the attention weights selected by the previous selection mechanism, we perform fusion normalization in the last step, so that the final channel attention map can be obtained. Here, we use the element-sum operation, which can minimize the parameters and improve the calculation efficiency. As shown in Figure 5, after using the attention method proposed in this paper, the information of the shallow feature map is more abundant, and the positioning and category semantic information is rich. As shown in Figure 6, after using the method designed in this paper, we can observe the visual information in the shallow feature map of the network.

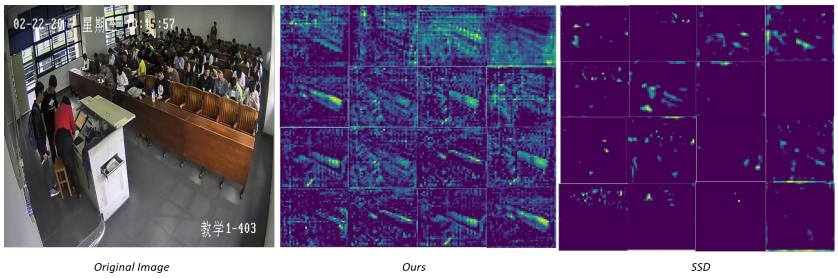

Fig. 5: Visual comparison feature map.

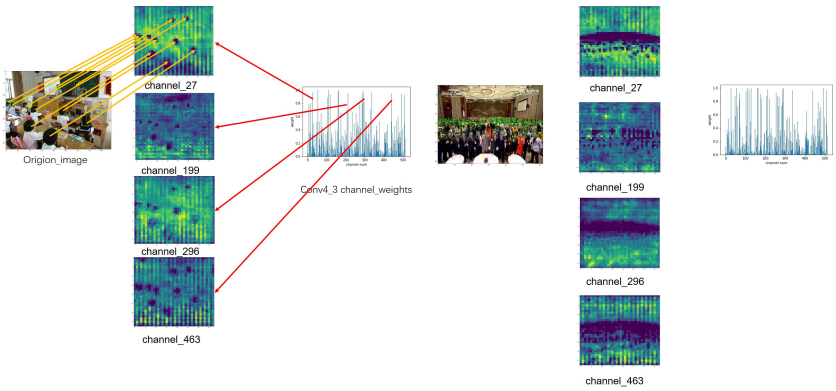

Fig. 6: Visualization results in the shallow feature map of the network.

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[20] pre-trained parameters on MSCOCO [15] 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. This is a large-scale head detection dataset,which follows the standard of Pascal VOC, including 4,405 images labeld with 111,251 heads. This dataset consists of two parts. PartA includes 2,000 images sampled from monitor videos of classrooms in an university with 67,321 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.

3.3 Experimental Comparisons

SCUT-HEAD. Table 2 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 3 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	Con-local	SSD	R-FCN	FRN	Ours
AP	44.5	80.2	84.8	88.1	**90.0**

Table 1: Comparison of other methods and our method on Brainwash Dataset

Brainwash. We also compare our method on Brainwash dataset in Table 1. 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 [10].

Method	PartA			PartB		
	P	R	F1	P	R	F1
Faster-RCNN[12]	0. 86	0. 78	0. 82	0. 87	0. 81	0. 84
YOLOv3[11]	0. 91	0. 89	0. 89	0. 74	0. 67	0. 70
SSD	0. 87	0. 68	0. 76	0. 80	0. 66	0. 72
R-FCN(ResNet-50)[1]	0. 87	0. 78	0. 82	0. 90	0. 82	0. 86
R-FCN+FRN[10]	0. 89	0. 83	0. 86	0. 92	0. 84	0. 88
DSCA-Net(proposed)	**0. 93**	**0. 91**	**0. 92**	**0. 95**	**0. 93**	**0. 95**

Table 2: 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. 23**	**0. 21**	**0. 22**	**0. 87**	**0. 84**	**0. 85**

Table 3: Comparison of other methods and our methods on small head detection

3.4 Qualitative Results

We show some visualization results of our method and other methods, as shown in Figure 7, Compared with the visualization results of other methods, it can be seen that the HD-Net proposed in this paper solves the problems of multi-scale and similar object and environmental characteristics.

4 Conclusion

In this paper, we propose a novel single-stage indoor head detection network called DSCA-Net, which is mainly used to detect indoor populations, and obtain the final population count statistics based on the test results. DSCA-Net consists of two modules: Dual Stream information Module (DSM) and Channel-Attention Module (CAM). DSM is used to solve the problem of multi-scale object, while CAM is used to solve the problem of similar object and background features. Our method achieves an F1 score of 0. 91 and a recall rate of 0. 92 on a standard datasets. The method in this paper is flexible and simple, and can also be migrated to other object detection tasks.

References

1. 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:*

Fig. 7: Qualitative results of HD-Net on the validation set of the SECUT-HEAD dataset.

Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 379–387, 2016.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409. 1556, 2015.
14. 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.
15. 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.
16. 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.
17. 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.
18. H. Zhang, I. J. Goodfellow, D. N. Metaxas, and A. Odena. Self-attention generative adversarial networks. *CoRR*, abs/1805. 08318, 2018.

19. 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.
20. 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.
21. 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.
22. 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.