Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Channel and Spatial Attention Fusion Module for Detection

Jun Yan
Zhejiang Normal University
Rong Zhu
Jiaxing Nanhu University
Bin Chen
Jiaxing University
Huiying Xu
xhy@zjnu.edu.cn

Zhejiang Normal University
Xinzhong Zhu
Zhejiang Normal University

Research Article

Keywords: Object Detection, Computer Vision, Feature Enhance, Attention

Posted Date: April 18th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2804607/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

The importance of the attention mechanism in CV is growing, as it allows a neural network to focus more on what it should pay attention to. Channel attention and spatial attention are the two basic attention strategies now in use. Using one of them alone can enhance to some level, while combining the two is more beneficial, but it adds to the computational burden. We propose a Channel and Spatial Attention Fusion Module(CSAFM), besides the use of channel and spatial information, GroupNorm and reorganization operation are applied to enhance the ability of feature extraction and feature representation of feature maps, which can capture more details as the feature deepens. Only a few additional parameters are added while enhancing the network. It is a plug-and-play module. We validate our CSAFM through extensive experiments on PASCAL VOC2007 and MS COCO2017 datasets, which shows that our module is effective yet efficient.e.g., it has risen by 1.47%, 2.1%, and 1.59%, respectively, under the frameworks of VGG16, ResNet50, and ResNet101 compared to baseline. Code is available at https://github.com/zero-yan/FasterRCNN.git.

1 Introduction

Convolutional neural networks and attention mechanisms in vision have been the subject of much research[1–4], with promising results in terms of improving the performance of the network in the area of computer vision. Attention mechanism[5– 7, 33–35] promotes the effect of CNN network from different aspects. CNN networks like VGGNet[8], GoogLeNet[9], and Residual-Style Networks[10–13] primarily enhance the network detection capacity by widening and deepening network, whereas the attention mechanism enhances the network performance by enhancing the expression of features and concentrating on crucial feature information and specifics. Adopting the attention mechanism will only marginally or barely increase the network's parameters and computation while concurrently improving performance in comparison to a network with more depth and breadth.

The purpose of the attention mechanism is to increase the percentage of valuable information while decreasing the proportion of useless information, which indicates that attention will not only tell us what and where to pay attention, but also enhance the representation of features. In computer vision, there are primarily two types of attention mechanisms that are applied: channel attention and spatial attention. The primary goal of channel attention is to locate and collect key channels and global information, whereas the goal of spatial attention is to locate and capture essential spatial and regional information. SENet[1] proposes the SE module (squeeze and excitation), which primary focus on the channel attention module, that can effectively deals with channel relationships. By displaying the dependency between modeling channels, it weights in the channel direction and utilizes global information to selectively highlight valid features and supress invalid features. However, SENet's[1] adaptability is weak, and as the network grows, a significant number of new parameters will be introduced, slowing down the model, and it ignores the spatial information. CBAM[2] and GCNet[3] start from the two scopes of channel and spatial, introduce two dimensions of spatial attention and channel attention, and realize the attention structure of integrating space and channel. Concatenating channel attention and spatial

attention along two separate dimensions is what CBAM[2] works (channel and space). It simultaneously employs global average pooling and maximum pooling to create an attention map, which is then multiplied by the input feature map to improve the capability of feature representation. GCNet[3] integrates spatial attention and channel attention through a global context modeling module to capture long-range dependencies, which is of great significance in deep neural networks. However, they either struggle to transmit the information across the two channels or should do a substantial portion of calculation. DANet[32] utilizes two attention modules(position attention and channel attention) to more effectively capture global dependencies and long-range dependencies. However, matrix computation makes the algorithm more complex and the model is less robust. ECANet[14] demonstrates that the channel attention mechanism will have negative consequences from the SENet's[1] dimensionality reduction. The whole connection layer is removed and the cross-channel information interaction is completed by 1-D convolution in the proposed local cross-channel interaction technique without dimensionality reduction. Wang X et al.[15] introduced a Non-Local Neural Network model, to efficiently capture the dependency between each element in the sequence. In object detection, the sequence may include various places of a single image (i.e. spatial sequence). But if the feature map is too large, the quantity of calculation will be excessive. The fundamental idea behind CCNet[16] is to limit the similarity computation between each Non-Local Neural Network[15] and all other locations to only the same row and column. This considerably reduces computation. However, they were less efficient because they did not fully exploit the relationship between spatial and channel attention.

In this paper, we introduced an attention module, named "channel-spatial Attention Module", which groups the dimensions of channel into sub-feature. To obtain channel attention, the global average and maximum average pooling operations are performed for each sub-feature. GroupNorm is applied to obtain spatial attention. Finally, the "channel shuffle"[17] unit is utilized to merge channel and spatial attention in order to allow cross-group information flow along the channel dimension. As a result, our model may enhance network identification performance with very few additional parameters. We demonstrate the wide application of CSAFM by verifying performance improvements in object detection on the MS COCO and PASCAL VOC

2007 datasets.

Contribution. The main contribution in this paper are summarized as follows:

1)We proposed a lightweight yet effective attention module(CSAFM) that can be widely applied to enhance representation power of CNNs.

2)Extensive ablation studies are used to prove the efficiency of our attention module. By plugging in our light-weight module, we establish that performance of different networks is considerably improved on numerous benchmarks (MS COCO, and VOC 2007).

2 Related Works

Object Detection. One of the main issues in the subject of computer vision is the task of object detection, which is to identify all interesting items (objects) in image and establish their categories and locations. Object detection has long been the most difficult topic in the field of computer vision because of the varying look, form, and posture of diverse objects, as well as the interference of imaging light, occlusion, and other factors. Deep learning-based object detection algorithms are often separated into two categories: two stage detectors and one stage detectors. Two stage detectors first generate the region, which is called region proposal (RP for short, a pre-selection box that may contain the object to be tested), and then make predictions for these proposals. Faster RCNN[25], as a representative of two stage detectors, achieved 73.2% of the tested map in the PASCAL VOC2007 test set. A region proposal network (RPN) was suggested as a replacement for selective search, to create the regions to be detected and suggested, and then to predict and classify, which implements a full end-to-end CNN object detection model. Mask RCNN[21], SPPNet[22], R-FCN[23], etc, are other popular two-stage object detection algorithms.

One stage detectors take dense samples at different positions of image, and then use CNN network to extract features and directly classify and regression. YOLO[29] treats the object detection task directly as a regression problem, and combines the two stages of candidate area and detection into one, which greatly improves the time efficiency at the cost of reducing mAP. Different from YOLO's full connection detection, SSD[30] convolutes the convoluted feature map of each layer and then detects it. At the same time, it uses the anchor concept in Faster RCNN[25] for reference and introduces a priori box, which can get better detection effect on small pictures. Other one stage detectors include RetinaNet[31] and et.al.

Features processing. Table 2 below shows that the majority of attention processes generate more additional parameters as the network's depth increases. Our approach utilizes grouping, avgpool, maxpool, GroupNorm[19] and other procedures to address the aforementioned issues. AlexNet[18] introduced the concept of grouping, which later transformed into the modern grouping convolution. Without taking into account the weight term, the parameter amount for traditional convolution for the input feature map $F \in [C,H,W]$ is: $C_{in} \times k \times k \times C_{out}$, whereas grouping convolution divides feature F into G groups, each of which produces a feature map, for a total of G feature maps to be produced. The size of each group is $F \in [C/G,H,W]$ and the size of each group of a single convolution kernel is $k \times k \times C/G$. Then, $C_{in} \times k \times k \times C_{out}/G$ is the parameter amount for grouping convolution. In order to divide the feature map and achieve the goal of lowering extra parameters, our model leverages the concept of grouping. Downsampling is a function of pooling. The more overall feature information of the image or feature map may be retained while the calculation parameters are reduced and the calculation speed is increased when the feature map is huge, which is helpful for the subsequent classification tasks. Due to the vast amount of input image data for the tasks of object detection, image segmentation, and video classification, big batches are frequently unable to function correctly, and the batch size for these deep learning tasks is frequently only 1–4. GroupNorm[19] is a normalization method, when compared with BN (batch normalization), GroupNorm[19] has an error rate that is almost the same under big batch sizes but around 10% lower under small batch sizes, which has a greater impact on increasing the model's

detection accuracy. Finally, it is necessary to transmit the feature maps information amongst various groups. If they are not shared, the network's capacity to extract features will be limited, and the intended outcome won't be realized. We utilize the "channel shuffle"[17] operation to "reorganize" the features in order to enable feature communication for each group so that information can flow between them.

Attention Mechanism. In recent years, the attention model has been widely applied in a variety of deep learning applications. In tasks such as image processing, speech recognition, and natural language processing, the visual attention mechanism is a unique brain signal processing mechanism of human vision. Human vision can quickly scan a global image to identify the target area that requires attention, i.e. the general focus of attention, and then devote more attention resources to this area in order to obtain more detailed information about the target that requires attention, while suppressing other irrelevant information. SENet[1] modeles the relationship of channel-wise by using the squeeze operation to compress features and using two full connection layers to recover to the original dimension. ECANet[14] enhances [1] by proposing a method for adaptively determine the size of a 1-D convolution kernel to estimate the coverage of local cross-channel interaction. Non-Local Neural Network[15] directly captures distant dependencies by calculating the interaction between any two locations, while CCNet[16] decreases the amount of calculation by restricting the calculation of [15] to the same rows and columns. CBAM[2] and GCNet[3] achieve new results by combining channel and spatial attention.

3 Materials And Methods

3.1 CSAFM

Given an intermediate feature map $F \in R^{C \times H \times W}$ as input, where C, H, and W stand for channel number, spatial height, and width, respectively.In order to better obtain feature extraction information and feature representation, CSAFM first extracts the global features X from F, this step is to fuse the global feature information with the sub-feature X_k to make the feature information more sufficient, so that it can better exchange information and circulate when the channel and spatial attention are fused. Specifically, $X_{k1}, X_{k2} \in R^{C/2G \times H \times W}$. Then, using an attention module, we calculate the associated significance coefficient for each subfeature. Finally, we aggregate all the sub-features and make some processing to allow cross-group information flow along the channel dimension. The following describes the details of each attention module.

Channel Attention Module. The attention mechanism is derived on the human visual attention mechanism, which imitates people's focus on different objects, and the channel attention is intended to focus on 'what' is most significant when given an input image. For aggregating information, CBAM(Woo et al.[2]) empirically confirmed that adopting both max-pooled and average-pooled features enhances network representation power far more than utilizing each alone.

We first aggregate global information of a feature map by using both averagepooling and max-pooling operations, generating two different descriptors:$X_{avg}{}^{C}$ and , then forward to a scale network to produce

channel attention map(MLP) to produce channel attention map $M_c \in R^{C/2G \times 1 \times 1}$, which used for adjust X's shape to facilitate the subsequent element-wise summation with subfeature. After that, averagepooling and max-pooling operations is applied to both two branches of each subfeature X_k, then, we merge the output feature vectors using element-wiseummation as $x_c \in R^{C/2G \times 1 \times 1}$, which can be computed as:

$$M_c(X) = MLP(AvgPool(X)) + MLP(MaxPool(X)) \quad (1)$$

$$x = AvgPool(X_{k1} + X_{k2}) + MaxPool(X_{k1} + X_{k2}) + M_c(X) \quad (2)$$

Finally, a compact feature is also built to provide for accurate and adaptable selection assistance. This is achieved by a simple gating mechanism with sigmoid activation. After that, you can obtain the final output of channel attention by:

$$xc = \delta(Fc(x)) \cdot X_k1 = \delta(Wcx + bc) \cdot X_k1 \quad (3)$$

Where δ denotes the sigmoid functions, $W_c \in R^{C/2G \times 1 \times 1}$, and $b_c \in R^{C/2G \times 1 \times 1}$ are parameters used to scale and adjust x.

Spatial Attention Module. Spatial attention differs from channel attention in that it is focused on a specific location. Channel attention is focused on 'what' features in the feature map are more essential, whereas spatial attention is focused on 'where' the key feature information is located in the feature map. To compute the spatial attention, we apply GN[19] over two branches X_{k1}, X_{k2} to reduse extra computation burden, which makes the spatial information of the object can be better applied and provide more detail information for the feature extraction network. The spatial attention is computed as:

$$xs = \delta(Ws \cdot (GN(X_k2) + GN(X_k1)) + bs) \cdot X_k2 \quad (4)$$

Where W_s and b_s are parameters with the following shape $R^{C/2G \times 1 \times 1}$.Then the two branches are then joined together to make the number of channels equal to the number of inputs.

All of the sub-features are then combined after that. Finally, to facilitate crossgroup information flow along the channel dimension, we employ a "shuffle"[24] operator. The CSAFM module output has the same size as feature X. And to detect our attention model in experiments, we primarily utilize the Faster RCNN[25] structure as our detector along with VGG16, ResNet50, etc.The network's general structure is shown in Fig. 1.

3.2 Visualization and Interpretation

In order to demonstrate the efficacy and feature representation of the CSAFM model on CNNs more intuitivelywe sample 9 photographs from the PASCAL VOC2007 data set. Grad-CAM[26] is an visualization method which uses gradients in order to calculate the significance of the spatial locations in convolutional layers. We use GradCAM[26] to visualize their heatmaps on CSAFM-ResNet50. For comparison, we also draw their heatmaps of ResNet50. As illustrated in Fig. 2, our proposed model

allows the classification model to focus on more relevant regions with more object information, which means the CSAFM model can effectively improve the classification accuracy.

Fig. 2 Grad-CAM[26] visualization results. We compare the visualization results of CSAFMintegrated network (ResNet50 + CSAFM) with baseline (ResNet50). The Grad-CAM[26] visualization is calculated for the last convolutional outputs. We sampled 9 images at random from the PASCAL VOC2007 dataset and transformed them into heatmaps for comparison.

4 Experiments

We evaluate CSAFM on PASCAL VOC2007 and MS COCO for object detection. To make better apples-to-apples comparisons, we reproduced all of the analyzed networks in the PyTorch framework and reported our reproduced results throughout the experiments.

We first conduct comprehensive ablation studies to thoroughly analyze the efficiency of our final module. Then, we verify that CSAFM outperforms all the baselines without bells and whistles, indicating the universal applicability across multiple architectures and tasks. CSAFM can be smoothly integrated into any CNN architecture.

4.1 Ablation Studies

In this part, we experimentally demonstrate the efficacy of our design decision. We utilize the PASCAL VOC2007 dataset and ResNet50 as the basis architecture for this ablation investigation. The PASCAL VOC2007 dataset consists of 5011 images for training and 4952 test images for validation with 20 object classes. We train networks on the training set and report the mAP on the validation set. All experiments are conducted with exactly the same residual structure and hyper-parameters settings in [27] and [1].To get better results,the initialization of parameters in $F_{ma}(\cdot)$ are set to 0 and 1 for weights(W_c and W_s) and bias(b_c and b_s) respectively. All the architectures are trained from scratch by Adam with weight decay 5e-4, gamma 0.96 for 100 epochs, starting from the initial learning rate 1e-4. For the testing on the validation set, the shorter side of an input image is first resized to 256, and a center crop of 224 × 224 is used for evaluation.

In [2], they experimentally verify that adopting both poolings features enables finer attention inference. Therefore, we further evaluated how channel attention, spatial attention, and fused attention affected our capacity to recognize networks. We utilized ResNet50 as the baseline and adopted SE[1], ECANet[14], and CSAFM-S (just adding channel attention) to compare channel attention. We employed CBAM[2], CSAFMC, and Non-Local[15] for space attention (both CBAM[2] and CSAFM only adding spatial attention). CBAM[2], GCNet[3] and our attention module are contrasted in fused attention. Table 1 demonstrates the ultimate outcomes. The chart makes it obvious that distinct space and channel attention are what causes the fused attention, and our attention model has produced the best outcomes.

Table 1
Comparsion of difference between baseline and different attention approaches. We observed that the best results were achieved by the fused channel and spatial attention module, and our model achieved the best results.

Backbone	Params	GFLOPs	mAP@.5
ResNet50(baseline)	27.15	517.44	77.46
ResNet50 + channel(SE[1])	29.56	518.39	78.64
ResNet50 + channel(ECANet[14])	27.15	517.44	**79.24**
ResNet50 + channel(CSAFM-C)	27.60	529.01	79.01
ResNet50 + spatial(CBAM[2])	27.15	517.46	78.31
ResNet50 + spatial(Non-Local[15])	65.56	71.79	**78.92**
ResNet50 + spatial(CSAFM-S)	27.60	529.01	78.70
ResNet50 + channel + spatial(GCNet[3])	29.58	518.61	78.73
ResNet50 + channel + spatial(CBAM[2])	31.95	518.40	79.10
ResNet50 + channel + spatial(CSAFM)	27.60	529.01	**79.56**

Experimental results with various approaches are shown in Table 1. We observe that the last method achieves the best outcome, comparing the mAP improvement from the baseline and other approaches. It is empirically demonstrated that the last strategy is an effective way to improve performance without introducing more learnable factors.

At the same time, we visualize the baseline and CSAFM feature maps and convert them to grayscale maps for comparison. As seen in the Fig. 3, it is obvious that the CSAFM model we presented handles details better, and with the depth of the feature map, our model has superior expression capacity than the baseline.

4.2 Experiments on PASCAL VOC

n Pacscal VOC 2007, we demonstrate that adding our attention approach to VGG16, ResNet50, ResNet101, and other backbones can significantly improve detail collection and network performance. The findings are summarized in Table 2. According to the findings, our solution performed the best in the same framework, whereas params barely increased.

We adopt Faster RCNN framework[25], which is one of widely used and comparison method on the object detection. For the experiment, we reproduce VGG16, ResNet50 and ResNet101 in our PyTorch platform in order to estimate performance

Input Image CSAFM CSAFM CSAFM CSAFM CSAFM improvement of CSAFM accurately and achieve 75.75%, 79.56% and 84.29% mAP@.5 respectively, which are higher than the original accuracy reported in the original papers. We train all the models on the set of PASCAL VOC 2007 train, and evaluate on the PASCAL VOC 2007 test set. The total number of training epochs is 100. We use a weight decay of 0.0005 and adopt StepLR with the hyper-parameter gamma of 0.96.

The networks with CSAFM considerably outperform all baselines, demonstrating that the CSAFM can generalize effectively on various models in the large-scale dataset. We gain better classification performance as a result of the comparison. It means that our proposed approach is powerful, and as shown in Fig. 3, our module can obtain more details from feature maps, resulting in superior outcomes and performance.

4.3 Object Detection On MS COCO

Table 3 clearly shows that our methods have increased ResNet50 and resent101 mAP@.5 by 1.02 and 1.13 percent, respectively. Additionally, other APs have improved their accuracy.

We train the current detectors on the COCO train2017 split and evaluate bounding box Average Precision (AP) for object detection. Microsoft COCO dataset[28] involves 11.8k training images (“2017 train”) and 5k validation images (“2017 val”). The average mAP over different IoU thresholds from 0.5 to 0.95 is used for evaluation. Because the pytorch framework we use reads data in TXT format, but the MS coco data set is in JSON format, we convert the JSON data set to TXT format via code, read it, and then divide the training and test sets into 9:1 ratios for training and testing. We adopt Faster R-CNN as our detection method and ResNet50 and ResNet101 as our baseline networks.

Table 2

Comparsion of different baseline and different attention approaches on PASCAL VOC2007 in terms of network parameters(Param.), giga floating point operations per second (GFLOPs) and mAP(in %). We adopt the Faster RCNN[25] detection framework and apply SE, CBAM, CSAFM to the detectors. All the baselines are positively improved by CSAFM with hardly any new parameters.

Backbone	Detector	Param.	GFLOPs	mAP@.5
VGG16	Faster RCNN[25]	121.73	172.71	74.28
VGG16 + SE[1]		121.82	172.52	75.44
VGG16 + CBAM[2]		123.14	172.53	75.63
VGG16 + CSAFM		121.73	172.52	**75.75**
ResNet18		11.86	365.39	68.02
ResNet18 + SE[1]		11.94	365.43	68.34
ResNet18 + CBAM[2]		12.03	365.44	68.55
ResNet18 + CSAFM		11.93	365.41	**68.69**
ResNet34		21.50	516.01	77.28
ResNet34 + SE[1]		21.65	516.07	78.19
ResNet34 + CBAM[2]		21.80	516.08	78.45
ResNet34 + CSAFM		21.62	516.03	**78.73**
ResNet50		27.15	517.44	77.46
ResNet50 + SE[1]		29.56	518.39	78.64
ResNet50 + CBAM[2]		31.95	518.40	79.10
ResNet50 + CSAFM		27.60	529.01	**79.56**
ResNet101		45.26	544.89	82.70
ResNet101 + SE[1]		49.82	545.84	83.47
ResNet101 + CBAM[2]		54.31	545.86	83.89
ResNet101 + CSAFM		46.11	556.46	**84.29**

Table 3
Object detection mAP(%) on COCO val2017. We adopt the Faster RCNN[25] detection framework and apply our module to the base networks. The best records are marked in **bold**.

Backbone	Detector	Param.	mAP@.5	AP50:95	AP_{50}	AP_{75}
ResNet50	Faster RCNN[25]	27.82	44.69	26.3	44.5	26.4
ResNet50 + CSAFM		28.51	**45.71**	**27.0**	**45.6**	**27.5**
ResNet101		46.00	46.74	28.2	46.6	27.9
ResNet101 + CSAFM		46.84	**47.87**	**29.5**	**47.8**	**29.1**

5 Conclusions

In this paper, we have proposed an attention module CSAFM, a lightweight and efficient module to improve and enhance the representation power of CNN networks. CSAFM splits the feature maps into multiple subfeatures. For each sub-features, all sub-features are aggregated after maxpool, avgpool, and GroupNorm operations in the channel and spatial attention module. Finally, we utilize a reorganization operator to enable information flowing across different subfeatures. Our module effectively refines intermediate features by learning what and where to stress or suppress. To verify its efficacy, we conducted extensive experiments on various models and confirmed that CSAFM outperforms all the baselines on different benchmark datasets: MS COCO and PASCAL VOC2007. In addition, we visualize heat map and gray-scale map how the module exactly infers given an input image, from the figures we observed that our module induces the network to focus on target object properly and can significantly improve the performance of various deep CNN architectures.

Declarations

Acknowledgments

This work was supported by the Natural Science Foundation of China (project no. 61976196), Zhejiang Provincial Natural Science Foundation of China (project no. LZ22F030003).

Ethical Approval

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Jun Yan wrote the main manuscript text. All authors reviewed the manuscript.

Funding

This work was supported by the Natural Science Foundation of China (project no. 61976196), Zhejiang Provincial Natural Science Foundation of China (project no. LZ22F030003).

Availability of data and materials

The datasets used are all publicly available.

References

1. Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132–7141.
2. Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3–19.
3. Cao Y, Xu J, Lin S, et al. Gcnet: Non-local networks meet squeeze-excitation networks and beyond[C]//Proceedings of the IEEE/CVF international conference on computer vision workshops. 2019: 0–0.
4. Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li,Wangmeng Zuo, and Qinghua Hu, “Eca-net: Efficient channel attention for deep convolutional neural networks,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020. 2020, pp. 11531–11539, IEEE.
5. Mnih V, Heess N, Graves A. Recurrent models of visual attention[J]. Advances in neural information processing systems, 2014, 27.
6. Ba J, Mnih V, Kavukcuoglu K. Multiple object recognition with visual attention[J]. arXiv preprint arXiv:1412.7755, 2014.
7. Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[J]. Advances in neural information processing systems, 2015, 28.
8. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
9. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1–9.
10. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770–778.
11. Zagoruyko S, Komodakis N. Wide residual networks[J]. arXiv preprint arXiv:1605.07146, 2016.
12. Xie S, Girshick R, Doll´ar P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1492–1500.

13. Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700–4708.
14. Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li,Wangmeng Zuo, and Qinghua Hu, “Eca-net: Efficient channel attention for deep convolutional neural networks,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020. 2020, pp. 11531–11539, IEEE.
15. Wang X, Girshick R, Gupta A, et al. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 77947803.
16. Huang Z, Wang X, Huang L, et al. Ccnet: Criss-cross attention for semantic segmentation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 603–612.
17. Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6848–6856.
18. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.
19. Wu Y, He K. Group normalization[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3–19.
20. Lin M, Chen Q, Yan S. Network in network[J]. arXiv preprint arXiv:1312.4400, 2013.
21. He K, Gkioxari G, Doll´ar P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961–2969.
22. Purkait P, Zhao C, Zach C. SPP-Net: Deep absolute pose regression with synthetic views[J]. arXiv preprint arXiv:1712.03452, 2017.
23. Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks[J]. Advances in neural information processing systems, 2016, 29.
24. Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 116–131.
25. Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.
26. Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE international conference on computer vision. 2017: 618–626.
27. He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]//European conference on computer vision. Springer, Cham, 2016: 630–645.
28. Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. Springer, Cham, 2014: 740–755.

29. Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779–788.
30. Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21–37.
31. Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980–2988.
32. Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3146–3154.
33. Liu Y, Shao Z, Hoffmann N. Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions[J]. arXiv preprint arXiv:2112.05561, 2021.
34. Zhang Q L, Yang Y B. Sa-net: Shuffle attention for deep convolutional neural networks[C]//ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 2235–2239.
35. Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713–13722.

Figures

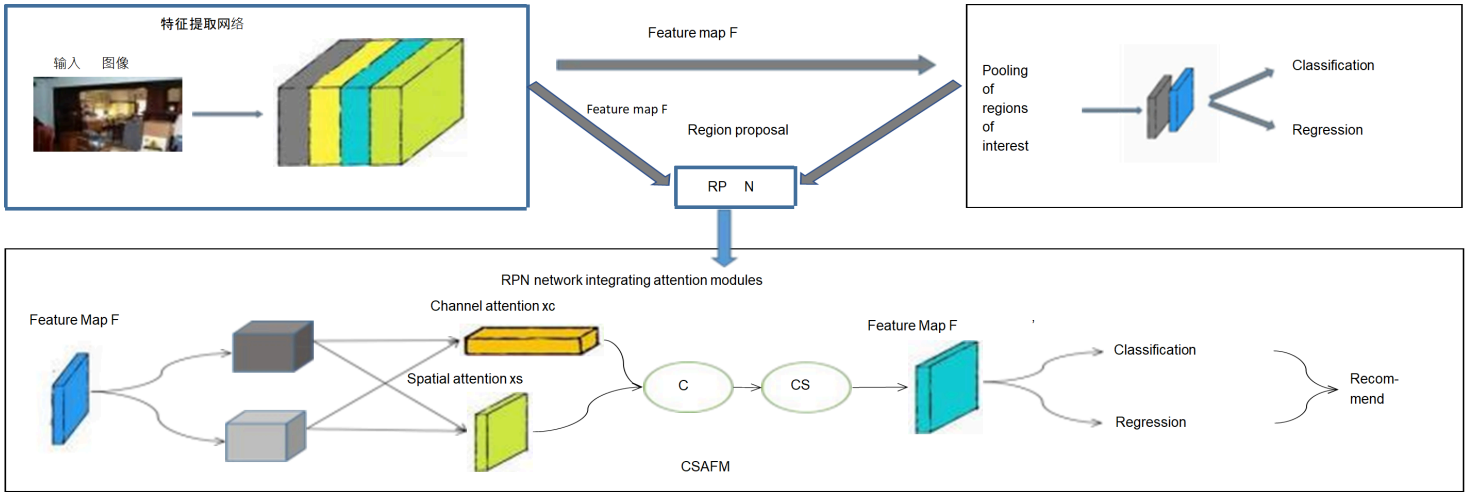


Figure 1

This is the architecture of Faster RCNN integrated with our attention model. Channel attention x_c and spatial attention x_s are obtained from the feature map F, and then channel shuffle is integrated to yield a new feature map F'.

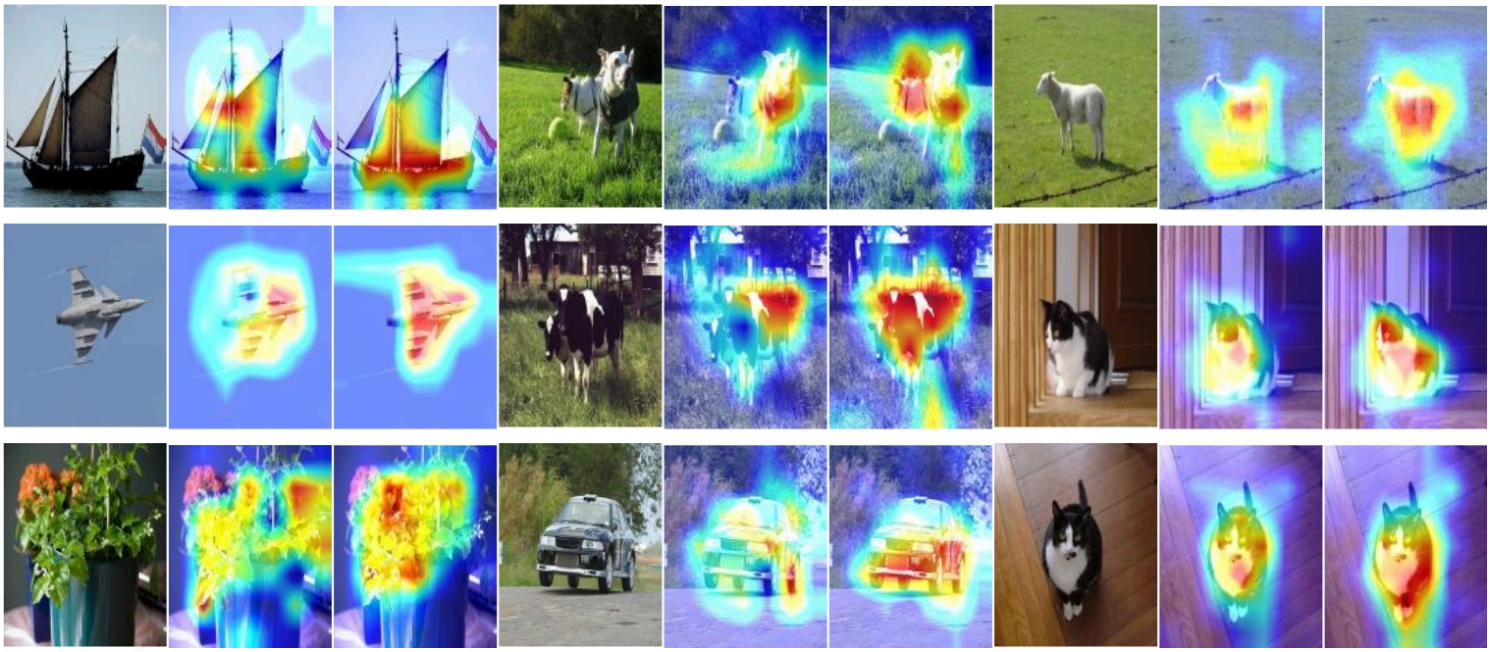


Figure 2

Figure legend not available with this version.

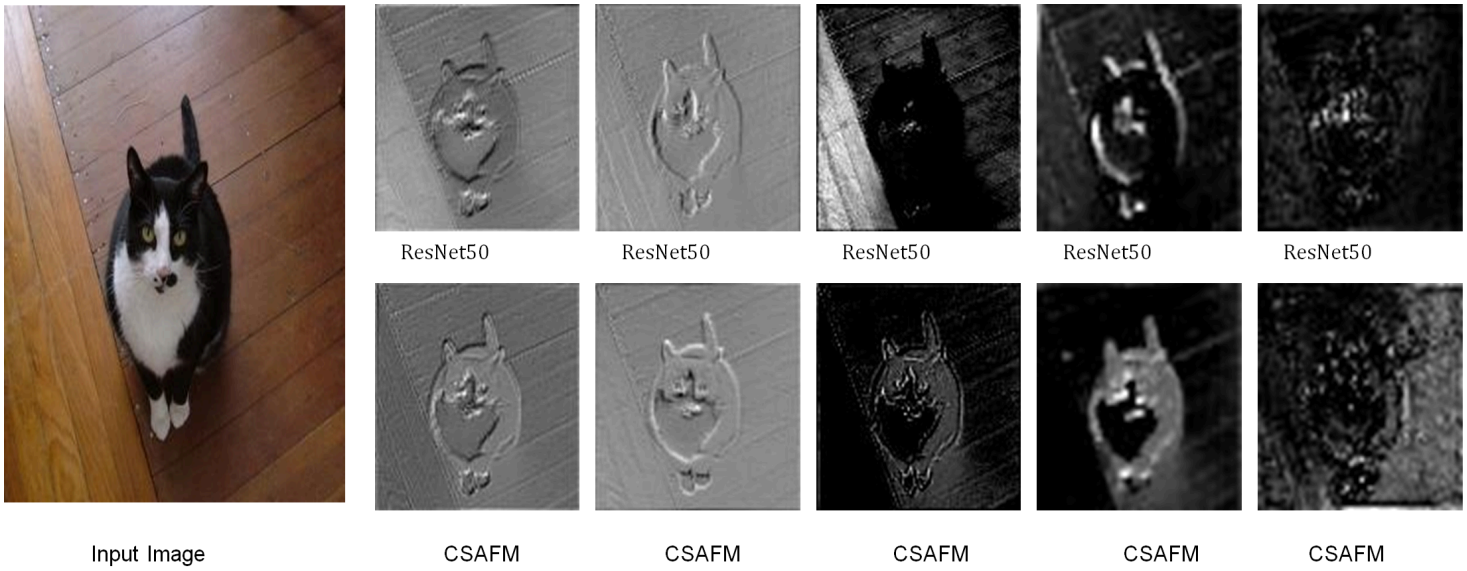


Figure 3

Convert baseline and CSAFM feature maps to grayscale maps for comparison. To evaluate the feature extraction capabilities of the two networks, we turn the feature maps from layer1 to layer5 into grayscale maps. The figure clearly shows that our suggested technique performs better in detail feature processing and can recover features with increasing depth.