Pattern Recognition Letters 154 (2022) 99–109

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Edge detection with attention: From global view to local focus

Huajun Liu[a,*], Zuyuan Yang[a], Haofeng Zhang[a], Cailing Wang[b]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[b] School of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China

ARTICLE INFO

Article history:
Received 2 August 2021
Revised 17 November 2021
Accepted 8 January 2022
Available online 11 January 2022

Edited by Jiwen Lu

Keywords:
Edge detection
Global attention
Local attention
Generalized zero shot learning
Pseudo distribution
Attribute similarity

ABSTRACT

Recently, edge detection models with convolutional neural networks (CNNs) have achieved significant advances, mostly by convolutional pyramid features and multi-path aggregation to generate accurate boundaries. However, it is still challenging for them when facing complex structure, weak context and dense boundaries. In this paper, we propose an Edge Attention Network (EdgeAtNet) from the viewpoint of attention mechanism. EdgeAtNet is derived from the richer convolutional features (RCF) basic architecture. On the low-level features, a global view attention block is inserted to the bottleneck to capture the long-range dependency of edge features, and on the high-level features, a local focus attention is designed for crisp boundary representation. Using ResNet101 as the backbone, we achieve state-of-the-art (SOTA) performance on several benchmarks. When evaluated on the well-known BSDS500 benchmark dataset, EdgeAtNet achieves the Optimal Dataset Scale (ODS) F-measure of 0.825. On the NYUDv2 and BIPED benchmark datasets, EdgeAtNet obtaines ODS F-measure of 0.764 and 0.868, respectively, and it outperforms the existing most methods.

1. Introduction

In an image, edges can be viewed as a set of contiguous pixel positions where abrupt changes of intensity occur, which usually corresponds to boundaries between world objects and background. Edge detection is a fundamental task with applications to semantic segmentation [1], image recognition, and object detection [2] etc.

Since the Sobel operator [3], many edge detectors [4] have been proposed and some techniques like Canny [5] are still widely used. Recently, CNN based edge detectors, for instance, DeepContour [6], HED [7], RCF [8,9], BDCN [10], and DexiNed [11], have significantly boosted the accuracy of edge detection (Fig. 1).

Some recent methods utilize hierarchical features with multi-scale representations by deep convolutional networks, and the VGG [12] or ResNet [13] are usually used as the backbone model. However, these methods (e.g., RCF [9], BDCN [10] and DexiNed [11] etc.) still suffer from the problem of limited receptive field caused by small convolution kernel size or downsampling operation. Therefore, these methods tend to fail in detecting long and thin boundaries in images, resulting in discontinuous edge detection.

In addition, some CNN models, such as HED [7] and RCF [9], benefit from parallel multi-path aggregation scheme, but also suffer from localization ambiguity in that they cannot focus on the object contour of interest very well, and they may generate much background clutter and blur edge even after the non-maximal suppression (NMS) post-processing.

It has been shown that false alarms in cluttered regions can be suppressed by seeking a sort of classification operation on the hierarchical deep features generated by large receptive fields [7,9,10]. Meanwhile, the attention gate mechanism [14], which has been introduced to medical image segmentation task, focused on local structures of varying shapes and sizes, and could learn to suppress irrelevant regions while highlighting local salient features. So in the edge detection task, it is desirable to find an efficient model with long range interaction and local contextual modeling simultaneously to represent the object contour precisely.

Inspired by these attention mechanisms, we propose an Edge Attention Network (EdgeAtNet) by combining a global view attention and a local focus attention to edge detection task. Specifically, (1) EdgeAtNet is based on the multi-path aggregation architecture on pyramid features, and a global view attention on low level features to capture long-range dependency and a local locus attention to generate more crispy semantic boundary. (2) To enrich the multi-scale contextual information of the image, a sequential scale-wise fusion module from coarse to fine is designed to fuse multi-scale features of each stage. (3) Finally, considering that localization ambiguity and edge expansion exist on high-level features, a

* Corresponding author.
E-mail addresses: liuhj@njust.edu.cn (H. Liu), yangzy@njust.edu.cn (Z. Yang), zhanghf@njust.edu.cn (H. Zhang), wangcl@njupt.edu.cn (C. Wang).

https://doi.org/10.1016/j.patrec.2022.01.006
0167-8655/

Fig. 1. The edge-maps predictions from the proposed model of wild images.

local focus attention (LFA) module on adjacent stages is designed to sharpen the edge maps.

2. Related works

2.1. Edge detection

As one of the most fundamental problems in computer vision, edge detection has been extensively studied for several decades. Early methods mainly focus on the utilization of intensity and color gradients, such as Canny [5]. However, the traditional methods are usually not accurate enough for wild images. To this end, feature learning based methods, such as Pb [15], gPb [16], and SE [17], usually employ sophisticated learning schemes to predict edge strength with low-level features such as intensity, gradient, and texture. Although these methods are shown to be promising in some cases, these handcrafted features are limited for semantically meaningful edge detection.

Many deep edge detectors have been introduced recently. For example, Xie and Tu [7] developed an efficient and accurate edge detector, HED, which is based on hierarchical pyramid features from VGG16 and proposed a holistically nested architecture to connect their side output layers. Moreover, Liu et al. [9] use relaxed labels generated by bottom-up edges to guide the training process. He et al. [10] proposed a bi-directional cascade structure for edge detection. Although these CNN-based models have pushed forward the state of the arts to some extent, they all turn out to be lacking in our view because they still have some shortcomings in wide-range edge structure capture and semantic crisp edge representation.

2.2. Multi scale representation learning

Extraction and fusion of multi-scale features are fundamental and critical for many vision tasks. Multi-scale representations can be constructed from multiple re-scaled images by concatenating features from different scale or using the output from one scale as the input to the next scale. Recently, innovative works DeepLab [18] and PSPNet [19] use dilated convolutions [18,19] and pooling to achieve multi-scale feature learning in image segmentation.

Like other image patterns, edges vary dramatically in scales. Ren's work [20] shows that considering multi-scale cues does improve performance of edge detection. For instance, HED [7] constructed multi-scale features learning for edge detection. Inspired by HED, RCF [8] fused features from bottom-up convolutional features to generate coarse-to-fine representation. BDCN [10] used a bi-directional cascade structure on multi scale convolution features for edge detection.

Multi-scale representation as a practical technique for edge detection has been verified in many models. Different with the parallel pipeline of multi-side fusion in HED [7], RCF [8] and BDCN [10], we further developed a sequential scale-wise fusion module from coarse to fine to improve the sharpness of edge.

2.3. Attention mechanism

Attention mechanisms have been introduced into many vision tasks to address the weakness of standard convolutions. For instance, Squeeze-and-Excitation [21], Gather-Excitep [22] and GC-Net [23] reweigh feature channels using signals aggregated from global context modeling for image classification and object detection tasks. Meanwhile, bottleneck attention module (BAM) [24] and dual attention network (DAN) [25] refine convolutional features independently in the channel and spatial dimensions for semantic segmentation task.

Global attention. Self-attention [26,27] and decomposed attention [24,28] have emerged as a recent advance to capture long range interactions, but has mostly been applied to sequence modeling and generative modeling tasks. Woo et al. [28] Park et al. [24] proposed decomposed global attention on channel and spatial dimension has been applied to image classification and object detection. These self-attentions and decomposed attentions are with global receptive field, but there are few reports on global-attention mechanism on edge detection.

Local Attention. Attention gate (AG) is commonly used in natural image analysis, knowledge graphs, image captioning [29] and machine translation [30] tasks. In the area of medical image analysis, spatial attention [14] or multi scale attention on U-Net [31] use attention gate mechanisms to combine local features with their corresponding global dependencies, explicitly model the dependencies between channels and use multi-scale predictive fusion to

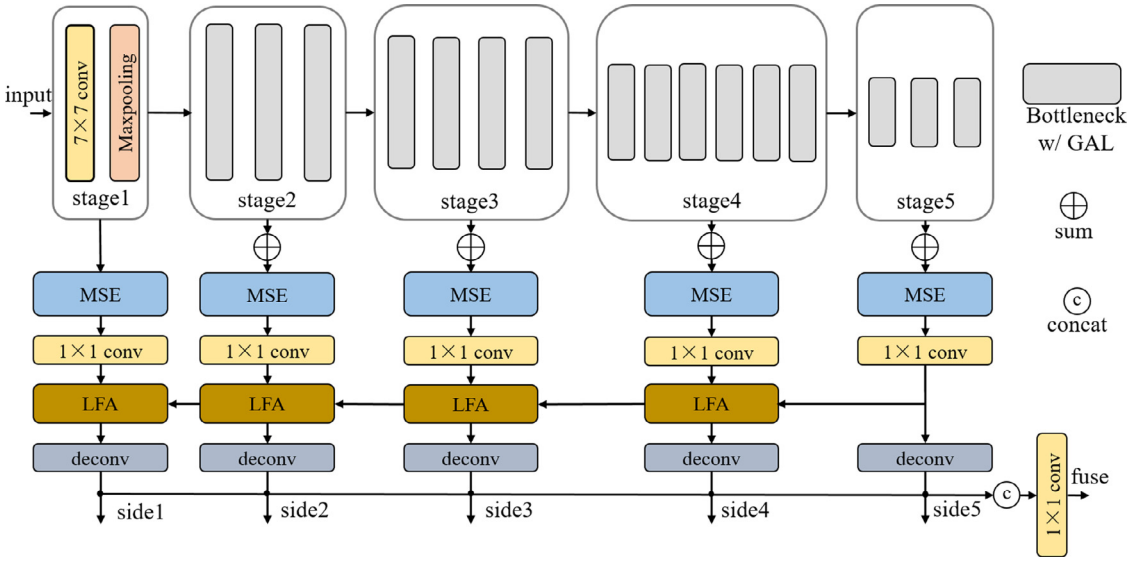


Fig. 2. Architecture of EdgeAtNet.

utilize different scale information in segmentation task. Oktay et al. [32] propose a soft attention mechanism to softly weight between the encoder features and decoder features at each pixel location. These soft attention or attention gate methods are operated on a local receptive field, up to now seldom research applies local attention for edge detection task.

3. Our EdgeAtNet network

3.1. Overview

Our work is derived from the RCF basic architecture, and it uses ResNet as the backbone and is built with sequential scale-wise multi feature aggregation on convolutional pyramid features. The architecture is shown as Fig. 2. We made some minor adjustments to the backbone. The stride of the 3×3 convolution kernel of the bottleneck in the first and last stages is set to 1, so the network has 3 times downsampling operation, and the resolution of the last stage is reduced to 1/8 of the original, which can maintain the details of the image. Since most salient edge features are both long and crisp, the receptive field of 3×3 stacked convolution cannot cover the long structure entirely. In our proposed EdgeAtNet, a decomposed global attention layer (GAL) is inserted to each bottleneck to enlarge receptive field to capture long range structure on low-level features. To enrich the multi-scale contextual information of the image, a sequential scale-wise fusion module is designed to refine the salient edge features scale by scale. Considering that localization ambiguity and edge expansion exist on high-level features, a local focus attention (LFA) module on adjacent layers is designed to sharpen the edge maps to generate more crispy boundary.

3.2. Long range capture with global attention layer

Global attention layer derives from the CBAM block [28], which is an efficient decomposed global attention composed of 'Avg Pooling' and 'Max Pooling' operations on channel dimension and spatial dimension respectively. Attention maps on channel and spatial branches are connected in series, followed by a Sigmoid activate function. Our proposed global attention layer upgrades the CBAM by a dynamic fusion strategy on channel combination, which is shown in Fig. 3(a).

Let X, X_1 and $X_2 \in \mathbb{R}^{C\times W\times H}$ be an input tensor, intermediate tensor and output tensor, where W and H represent the width and height of input tensor and C denotes their channel dimension.

On its channel dimension, two feature maps $X^{ch}_{avg} \in \mathbb{R}^{C\times 1\times 1}$ and $X^{ch}_{max} \in \mathbb{R}^{C\times 1\times 1}$ can be obtained by an adaptive average pooling (shown as $AvgPool_{ch}$) and an adaptive maximum pooling (shown as $MaxPool_{ch}$) operations on channels by compressing spatial dimensions. The $AvgPool_{ch}$ is an AdaptiveAvgPool2d operation with parameter 1, which compressed the 2D spatial domain and extracted the channel features of background after noise filtering, while the $MaxPool_{ch}$ is an AdaptiveMaxPool2d operation with parameter 1, which compressed the 2D spatial domain and excited salient edge features with rich details. Both of them operate on the global receptive field, which help to capture the long range structure of object edge. We upgraded the original CBAM block in a flexible way by considering how to combine these branches with a dynamic fusion strategy. Specifically, these two feature maps are dynamically fused by two parallel Conv 1×1 (shown as 'dConv' in Fig. 3(a)) for channel reassemble from $c \to c/r$ and from $c/r \to c$, but not by a fully connected layer. Then channel attention mask M_{ch} can be obtained through a Sigmoid activation function. Finally, the channel attention mask of M_{ch} and the input tensor X are multiplied element-wisely to obtain the feature map X_1 with the channel attention.

$$X^{ch}_{avg} = \mathrm{AvgPool}_{\mathrm{ch}}(X) \tag{1}$$

$$X^{ch}_{max} = \mathrm{AvgPool}_{\mathrm{ch}}(X) \tag{2}$$

$$M_{ch} = \sigma\left(\otimes_{1\times 1}\left(X^{ch}_{\mathrm{avg}}\right) + \otimes_{1\times 1}\left(X^{ch}_{\mathrm{max}}\right)\right) \tag{3}$$

$$X_1 = M_{ch} \cdot X \tag{4}$$

where $AvgPool_{ch}$ and $MaxPool_{ch}$ represent adaptive average pooling and adaptive maximum pooling operations on spatial dimension, respectively, $\otimes_{1\times 1}$ represents a dynamic convolution (denoted as 'dConv' in Fig. 3(a)) with a ReLU operation between 2 $1 \times 1 convolutions$, and $\sigma(\cdot)$ represents the output of the Sigmoid activation function.

On the spatial dimension, the global average pooling (shown as $AvgPool_{sp}$) and global maximum pooling (shown as $MaxPool_{sp}$) operations are performed on feature map X_1 along the channel dimension, respectively, to obtain $X^{sp}_{avg} \in \mathbb{R}^{1\times H\times W}$ and $X^{sp}_{max} \in \mathbb{R}^{1\times H\times W}$ by tensor mean and tensor maximum operation respectively. Then

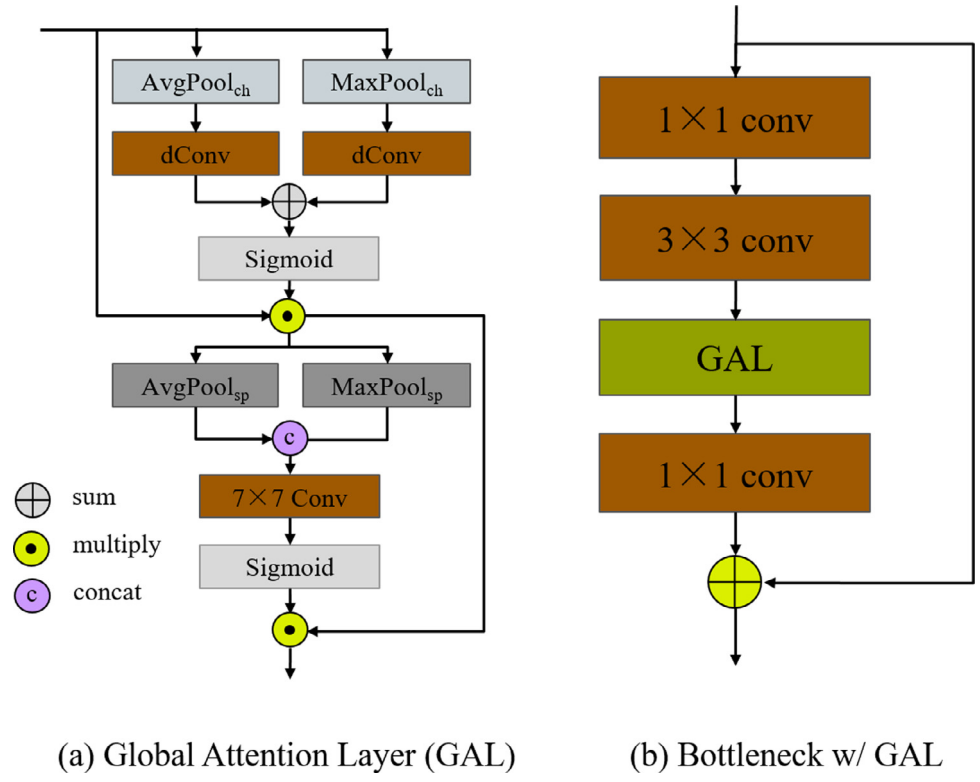


Fig. 3. GAL block and bottleneck.

these two features are concatenated on the channel dimension, and followed by a 7×7 convolutional layer and a Sigmoid activation function to get the spatial attention mask X_{sp}. Finally, the spatial attention mask M_{sp} and the tensor X_1 from the channel attention part are multiplied to obtain the feature map with both channel and spatial attention.

$$X_{avg}^{sp} = \text{AvgPool}_{\text{ch}}(X_1) \tag{5}$$

$$X_{\max}^{sp} = \text{AvgPool}_{\text{ch}}(X_1) \tag{6}$$

$$M_{sp} = \sigma\left(\otimes_{7\times7}\left(c\left(X_{\text{avg}}^{sp}, X_{\max}^{sp}\right)\right)\right) \tag{7}$$

$$X_2 = M_{sp} \cdot X_1 \tag{8}$$

where $c(\cdot)$ denotes tensor concatenation operation; $\otimes_{7\times7}$ denotes the convolution with kernel size of 7×7.

The global attention layer works as a plug-in inserted to the bottleneck of ResNet between the Conv 3×3 and the last Conv 1×1. The updated bottleneck is shown in Fig. 3(b).

The global attention maps obtained from test images of each stage are visualised as Fig. 4. We commonly observe that attention maps from low-level features have a uniform distribution and pass features at all locations. Additionally, at denser stages attention maps provide a rough outline of object of interest, which are gradually refined at finer resolutions.

3.3. Crispy boundary generation with local focus attention

High-level features often get more meaningful semantic features after many layers of convolution operations, such as the category information of the object, while low-level features tend to get the position information or edge contour information of some objects.

Different with the parallel pipeline of multi-side fusion of HED, RCF and BDCN, we further developed a serialized scale-wise fusion module from coarse to fine to improve the sharpness of edge. To extract crisp edge, we propose Local Focus Attention (LFA) module, seen as Fig. 5. At each stage, upsampled denser stage feature F_{k+1} is concatenated with the next stage feature F_k to form a concat tensor F_{cat}, followed by a 1×1 Conv is done to compress the channel. Then a *Mask* is obtained through the Sigmoid activation function. With an element-wise multiplication between the local attention map *Mask* and the current stage feature map F_k, we can obtain the LFA enhanced edge feature maps F_{fuse}.

$$F_{cat} = c(F_k, upsample(F_{k+1})) \tag{9}$$

$$F_{fuse} = F_k \cdot \sigma\left(\otimes_{1\times1} F_{cat}\right) \tag{10}$$

The attention maps for local focus at each stage are visualised as Fig. 6, which is an attention coefficient maps from high-level features. Around the semantic edge, there is a stronger response. From the output features of different scales, we can see a fine to coarse semantic edge response, which can be used to refine more crisp boundary.

3.4. Multi scale enhancement

Inspired by the dilated convolution to enlarge receptive field for semantic segmentation in PSP [18] for edge detection in BDCN [10]. A Multi-Scale Enhancement (MSE) module is inserted into each stage to enrich the multi-scale representations. For a feature map $x \in R^{H\times W}$ convolved with a dilated filter $x \in R^{h\times w}$, the output at location (i, j) is

$$\mathbf{y}_{ij} = \sum_{m,n}^{h,w} \mathbf{x}_{[i+r\cdot m, j+r\cdot n]} \cdot \mathbf{w}_{[m,n]} \tag{11}$$

where r is the dilation rate, indicating the stride for sampling input feature map.

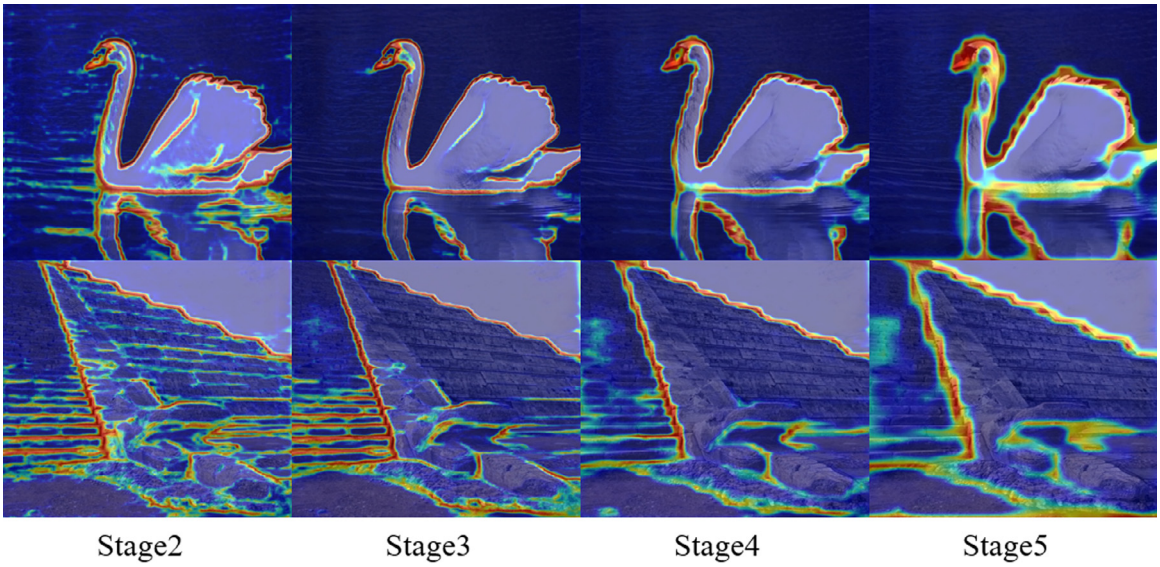


Fig. 4. Attention map from bottleneck w/ GAL.

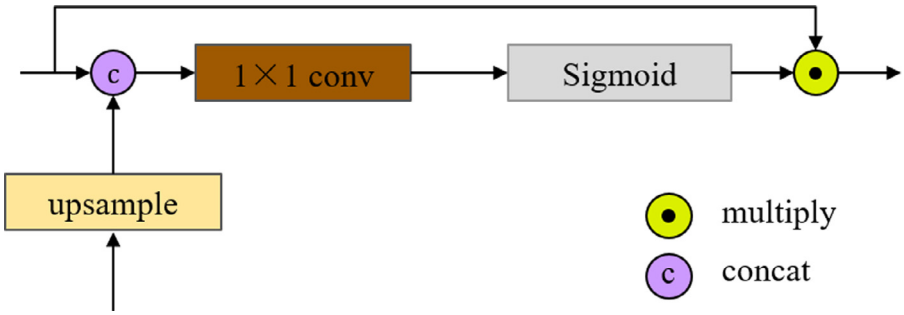


Fig. 5. Local Focus Attention (LFA) Block.

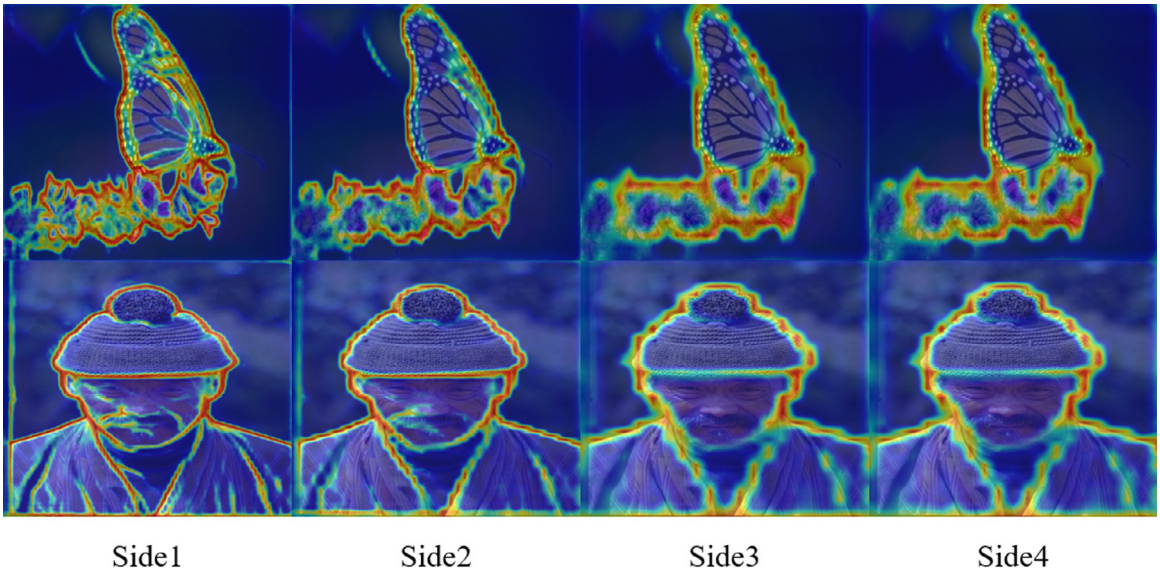


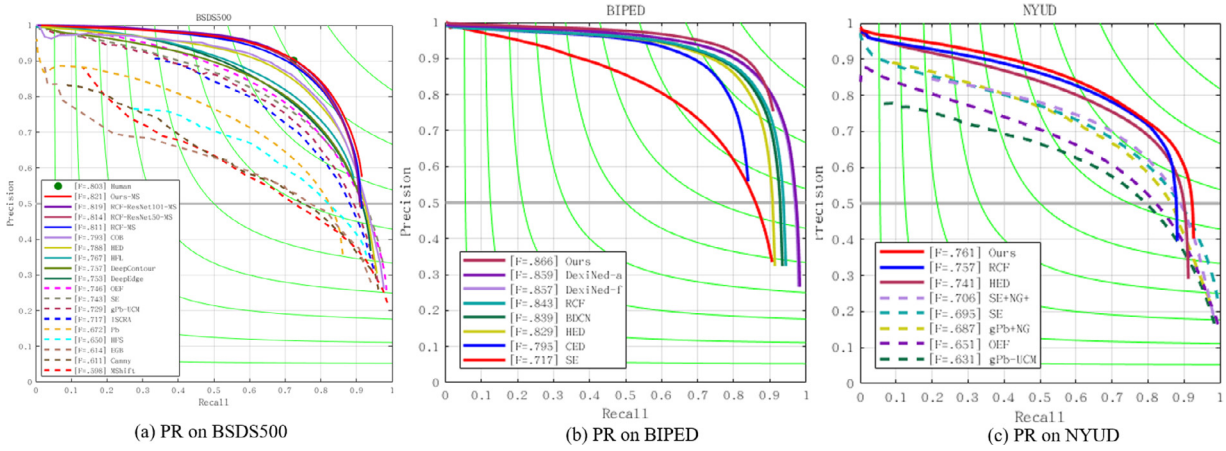
Fig. 6. Attention map from LFA.

3.5. Loss function

We make a modification on the balanced weight cross-entropy loss function ever used in the RCF network [9], where those pixels with probability higher than η are considered as positive samples, those pixels with probability 0 are considered as negative samples, and pixels in between as neglected. Similarly, the hyper-parameter η is used to balance the number of positive and negative samples. But in the modified loss function, a second hyper-parameter, e.g., 0.05 in our work, is used to cover those extreme weak negative

Table 1
EdgeAtNet for edge detection on the BSDS500 dataset.

Method	Backbone	ODS ↑	OIS ↑	AP ↑
MShift	-	0.598	0.645	0.497
Canny	-	0.611	0.676	0.520
EGB	-	0.614	0.658	0.564
HFS	-	0.650	0.688	0.201
Pb	-	0.672	0.695	0.652
ISCRA	-	0.717	0.752	0.770
gPb-UCM	-	0.729	0.755	0.745
SE	-	0.743	0.764	0.800
OEF	-	0.746	0.770	0.815
DeepEdge	-	0.753	0.772	0.807
DeepContour	-	0.757	0.776	0.790
HED	VGG	0.790	0.808	0.811
CED	-	0.803	0.820	0.871
RCF	VGG	0.806	0.823	0.839
RCF-MS	VGG	0.811	0.830	0.846
RCF	Res101	0.812	0.829	0.845
RCF-MS	Res50	0.814	0.833	**0.849**
RCF-MS	Res101	0.819	0.836	0.847
BDCN	VGG	0.806	0.826	0.847
BDCN-MS	VGG16	0.828	0.844	0.890
Ours				
EdgeAtNet	Res50	**0.821**	**0.837**	0.843
EdgeAtNet	Res101	**0.825**	**0.843**	0.848



(a) PR on BSDS500 (b) PR on BIPED (c) PR on NYUD

Fig. 7. PR curve on BSDS500.

pixels.

$$l(X_i;W)=\begin{cases}\alpha\log\left(1-P(X_i;W)\right) & 0\le y_i\le 0.05\\ 0 & 0.05<y_i\le\eta\\ \beta\log P(X_i;W) & y_i>\eta\end{cases}\tag{12}$$

$$\begin{cases}\alpha=\lambda\frac{|Y^+|}{|Y^+|+|Y^-|}\\ \beta=\frac{|Y^-|}{|Y^+|+|Y^-|}\end{cases}\tag{13}$$

where $|Y^+|$ and $|Y^-|$ represent the number of positive and negative samples, respectively. A parameter λ is used to balance the loss ratio of positive and negative samples. Let X_i represent the value of pixel i, and y_i is the edge probability of pixel i in the labeled image, and $P(X_i;W)$ represents the probability that the pixel is an edge, and W represents the weight of the model.

To reweigh each output side in training process, we weigh the loss on different side outputs, and increase the weights in the last two sides and the fusion side. The total loss function can be written as:

$$\mathrm{L(W)}=\sum_{i=1}^{n}\left(\sum_{k=1}^{5}S_{side}^{k}\cdot l\left(X_i^k;W\right)+S_{fuse}\cdot l\left(X_i^{\text{fuse}};W\right)\right)\tag{14}$$

where $S_{side}^{k}, k\in\{1,2,3,4,5\}$ represents the loss weight of the kth stage, S_{fuse} being the loss weight of the fusion layer, n being the total number of pixels in each sample, and k being the number of side outputs.

4. Implementation details

4.1. Data augmentation

On the BSDS500 benchmark, scaling (with ratio of 0.5, 1, and 1.5) and rotating 22.5° from 0° to 360° are common ways for data augmentation. On the NYUDv2 benchmark, rotating 90° and flipping horizontally on HHA and RGB samples are adopted. On the BIPED benchmark, samples are split up to half, and rotated by 15 different angles and cropped, and also flipped horizontally. In addition, two gamma corrections (0.3030, 0.6060) are applied.

4.2. Training/Testing parameters

As for loss function, On BSDS500, the η in the loss function is set to 0.3 because of multiple annotators. On BIPED and NYUDv2, η can be neglected for their binary labels. The λ of RGB and HHA

Table 2
EdgeAtNet for edge detection on the BIPED dataset.

Method	Backbone	ODS ↑	OIS ↑	AP ↑
SED	-	0.717	0.731	0.756
HED	VGG	0.829	0.847	0.869
CED	-	0.795	0.815	0.830
RCF	Res101	0.843	0.859	0.882
BDCN	VGG	0.839	0.854	0.887
DexiNed-f	-	0.857	0.861	0.805
DexiNed-a	-	0.859	0.867	**0.905**
Ours				
EdgeAtNet	Res50	**0.866**	**0.871**	0.885
EdgeAtNet	Res101	**0.868**	**0.875**	0.886

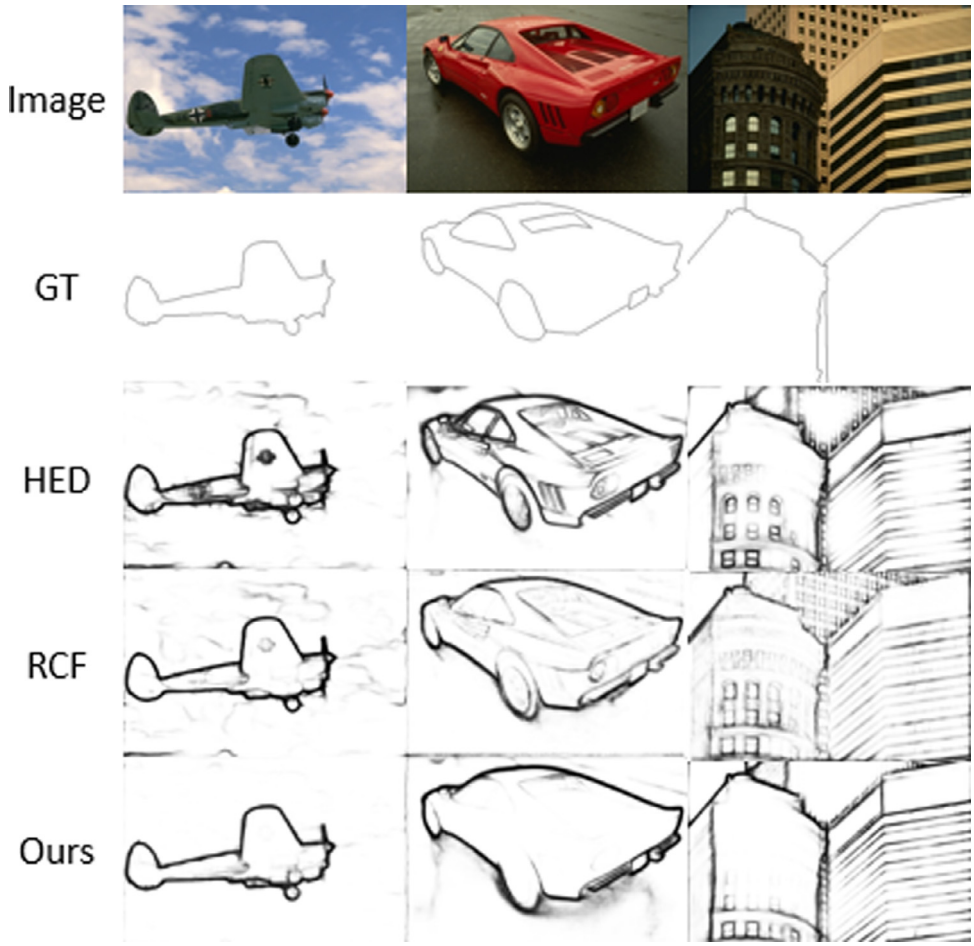


Fig. 8. Results on BSDS500.

are both set to 1.2 On NYUDv2, and is set to 1.1 on BIPED and BSDS500. The loss weights S^k_{side} of different sides are set as 0.5, 0.5, 0.5, 0.6, 0.6, respectively, and S_{fuse} set as 1.2.

During training, the stochastic gradient descent (SGD) is selected as the optimizer, and the momentum and weight decay are set to 0.9 and 0.0002, respectively. The initial learning rate lr_{init} is set to 0.001, it drops to 1/10 of the original after every 4 epochs. Batch size is set to 1, the weight is updated every 10 images using gradient accumulation, and the batch size increases in disguise. The backbone is initialized by pre-trained ResNet weights, and the remaining layers are initialized with a normal distribution $N(0, 0.1)$.

The post-processing scheme includes an NMS step and a morphology operation to obtain a thinned edge map. We use multi-scale test during the testing phase. Three scales of 0.5, 1.0, and 1.5 are used for inference.

5. Experiments and discussion

5.1. Datasets

We evaluate the proposed approach on 3 public datasets: BSDS500 [16], NYUD [33], and BIPED [11].

BSDS500 contains 200 images for training, 100 images for validation, and 200 images for testing. Each image is manually annotated by multiple annotators. The final groundtruth is the averaged annotations by the annotators.

NYUDv2 consists of 1449 pairs of aligned RGB and depth images. It is split into 381 training, 414 validation, and 654 testing images. This is a challenging dataset for indoor scene parsing and is also a commonly used benchmark for edge detection evaluation.

BIPED contains 250 outdoor images of 1280×720 pixels each. These images have been carefully annotated by experts on the

Table 3
EdgeAtNet for edge detection on the NYUDv2 dataset.

Method	Backbone	ODS ↑	OIS ↑	AP ↑
gPb-UCM	-	0.632	0.661	0.562
OEF	-	0.651	0.667	0.653
gPb+NG	-	0.687	0.716	0.629
SE	-	0.695	0.708	0.719
SE+NG+	-	0.706	0.734	0.549
HED-RGB	VGG	0.720	0.734	0.734
HED-HHA	VGG	0.682	0.695	0.702
HED-RGB-HHA	VGG	0.746	0.761	0.786
RCF-RGB	VGG	0.729	0.742	0.738
RCF-HHA	VGG	0.705	0.715	0.710
RCF-RGB-HHA	VGG	0.757	0.771	0.760
Ours				
EdgeAtNet-RGB	Res50	**0.746**	**0.760**	**0.752**
EdgeAtNet-HHA	Res50	**0.705**	**0.717**	0.700
EdgeAtNet-RGB-HHA	Res50	**0.761**	**0.779**	**0.787**
EdgeAtNet-RGB	Res101	**0.748**	**0.763**	**0.755**
EdgeAtNet-HHA	Res101	**0.709**	**0.723**	0.732
EdgeAtNet-RGB-HHA	Res101	**0.764**	**0.781**	**0.790**

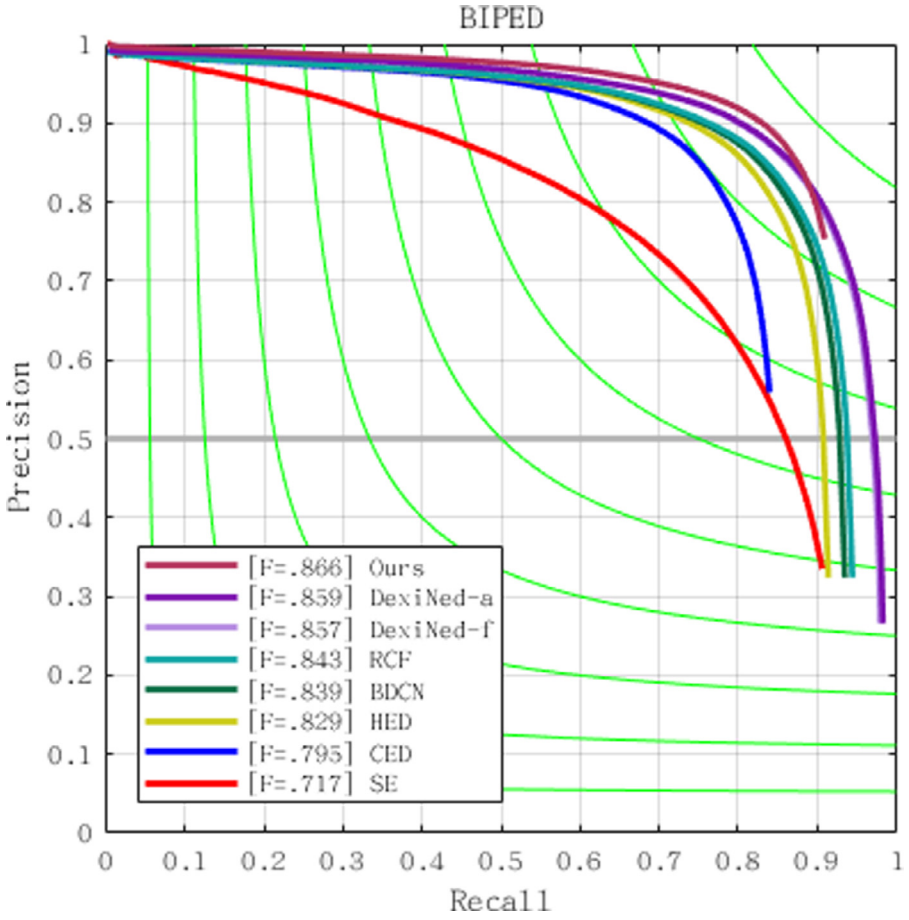


Fig. 9. PR curve on BIPED.

computer vision field, hence no redundancy has been considered. In spite of that, all results have been cross-checked in order to correct possible mistakes or wrong edges.

5.2. Performance metric

The performance metrics of *Precision* (abbre. as *PR*) and *Recall* (abbre. as *RE*) are calculated as $PR = \frac{TP}{TP+FP}$ and $RE = \frac{TP}{TP+FN}$ for binary classification tasks.

Given an edge probability map, a threshold is needed to produce the binary edge map. There are two choices to set this threshold. The first one is referred as *ODS* which employs a fixed threshold for all images in a dataset. The second is called *optimal image scale (OIS)* which selects an optimal threshold for each image. We report the F-measure $\left(\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}\right)$ of both ODS and OIS [17] in our experiments.

5.3. Comparison with SOTA methods

On the BSDS500 dataset, the comparison results are listed in Table 1, and the P-R curve is shown in Fig. 7. We use 'MS' to denote multi-scale testing for inference, and 3 different scales

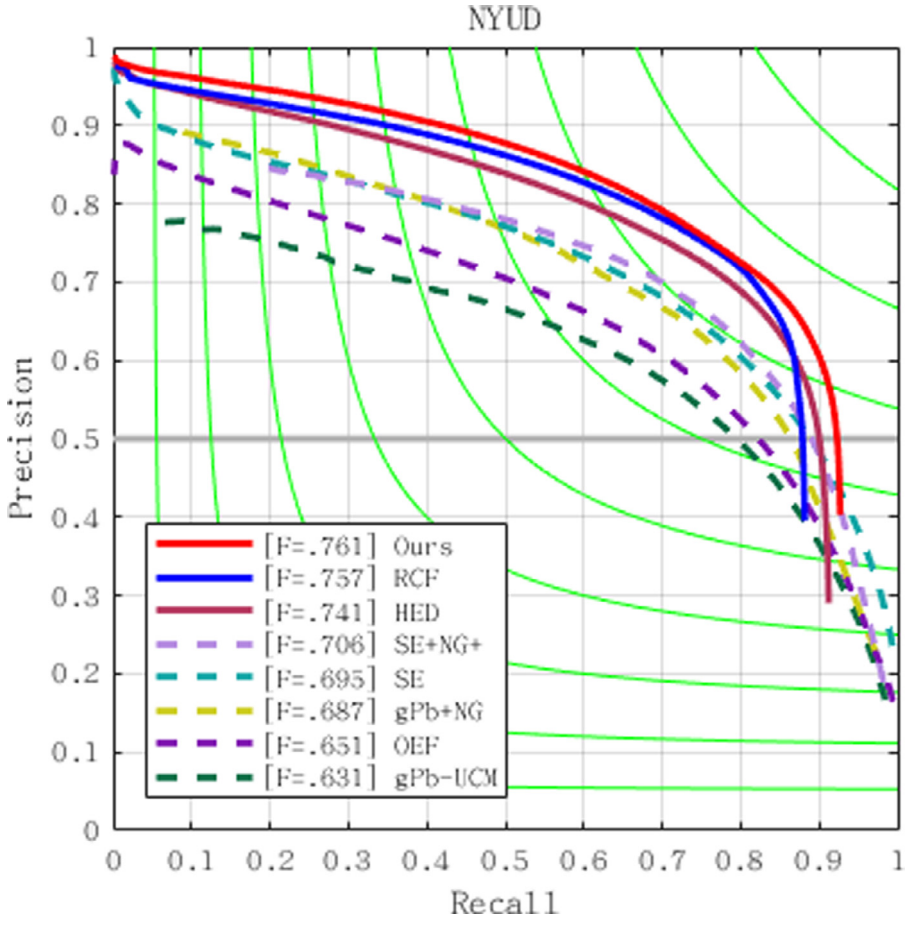


Fig. 10. PR curve on NYUDv2.

[0.5, 1.0, 1.5] are adopted. From the Table 1, EdgeAtNet-Res101 can increase by 0.6% on ODS and 0.7% on OIS compared with the RCF-Res101, and EdgeAtNet-Res50 increases by 0.7% on ODS and 0.4% on OIS compared with the RCF-Res50-MS. The edge results of our network, the HED and RCF models are shown in Fig. 8. It can be found that our prediction results are more closing to the object contour in the ground truth, but the edge prediction results by the HED and RCF models contain a lot of background clutter.

On the BIPED dataset, our results are compared with the mainstream models. The comparison results are listed in Table 2, and the P-R curve is shown in Fig. 9. Among them, DexiNed-f represents the result of the fused layer, and DexiNed-a represents the average result of all outputs. Compared with DexiNed, EdgeAtNet-Res101 increases by 0.9% on the ODS, and 0.8% on the OIS, and EdgeAtNet-Res50 increases by 0.7% on the ODS, and 0.4% on the OIS, respectively.

On the NYUDv2 dataset, our method is evaluated in three formats, namely RGB, HHA and RGB-HHA. The average of RGB and HHA output is regarded as the evaluation result of RGB-HHA. The compared results are shown in Table 3, it can be seen that the EdgeAtNet-Res101 increases the ODS by 1.9%, 0.4% and 0.7% on RGB, HHA and RGB-HHA respectively, and increases the OIS by 2.1%, 0.8% and 1.0% on RGB, HHA and RGB-HHA respectively. And the P-R curve is shown in Fig. 10. The inference results of our model on the NYUDv2 data set are shown in Fig. 11. It can be found that the prediction results generated by our method are more refined than the HED model and the RCF model, and are most similar to the ground truth.

Table 4
Ablation study on BSDS500.

CBAM	GAL	MSE	LFA	ODS ↑	OIS ↑
				0.813	0.832
	✓			0.814	0.833
		✓		0.817	0.834
			✓	0.815	0.833
	✓	✓		0.818	0.835
✓		✓	✓	0.819	0.836
	✓	✓	✓	0.821	0.837

5.4. Ablation study

To further check the gain of each module, e.g., GAL, LFA and MSE in our model, ablation study is done on BSDS500. Moreover, we further assessed the performance gain on the GAL module replacing the original CBAM block in a complete EdgeAtNet model. The experimental results are shown in Table 4. We select RCF-Res50 as the baseline. After the GAL module is inserted to the backbone, the gain on ODS and OIS is 0.1% and 0.1%, respectively, proving that the global attention is effective for edge representation. In addition, the MSE module can get gain of 0.4% and 0.2% on ODS and OIS, respectively, indicating that the model can fully integrate features of different scales. Moreover, the LFA module further achieves the gain of 0.3% and 0.2% on ODS and OIS, respectively, indicating that the local attention is important for crisp boundary representation. Experiments show that the GAL module obtained

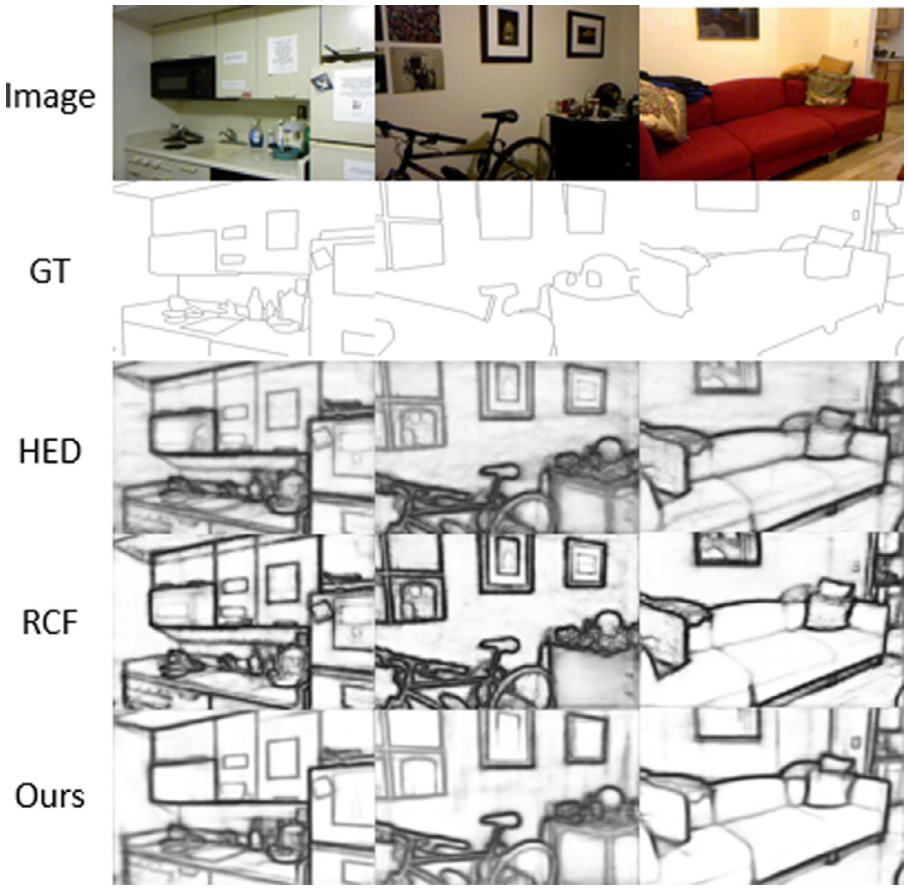


Fig. 11. Results on NYUDv2.

about 0.2% more F-measure gain on ODS on the BSDS500 dataset. Finally, we assessed the performance gain of a minor modification on the loss function definition in Section 3.5, which can get the F-measure gain about 0.05% on ODS on the BSDS500 dataset.

5.5. Efficiency analysis

The main operation of the new module, e.g., GAL and LFA in our EdgeAtNet is the element-wise multiplication. Assuming the pixel number of the input tensor X ($X \in \mathbb{R}^{C\times W\times H}$) is N, then $N = H \times W$, the complexity of GAL and LFA is $O(N)$, which is more efficient than most self-attention blocks [26,27]. However, the computational complexity of self-attention block is $O(N^2)$ because of the matrix Kronecker product and the Softmax operataions for long range interaction.

6. Conclusion

Our proposed EdgeAtNet is from the viewpoint of attention mechanism, which derives from the RCF basic architecture, on its low-level features, a global view attention block is inserted to the bottleneck to capture long-range dependency of edge features; on the high-level features, a local focus attention is designed for crisp boundary representation. Using ResNet101 as the backbone, we achieve SOTA performance on several benchmark datasets.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L.-C. Chen, J.T. Barron, G. Papandreou, K. Murphy, A.L. Yuille, Semantic image segmentation with task-specific edge detection using CNNs and a discriminatively trained domain transform, CVPR, 2016.
[2] V. Ferrari, L. Fevrier, F. Jurie, C. Schmid, Groups of adjacent contour segments for object detection, IEEE Trans. PAMI 30 (1) (2008) 36–51.
[3] N. Kanopoulos, N. Vasanthavada, R.L. Baker, Design of an image edge detection filter using the Sobel operator, IEEE J. Solid-State Circuits 23 (2) (1988) 358–367.
[4] V. Torre, T.A. Poggio, On edge detection, IEEE Trans. PAMI (2) (1986) 147–163.
[5] J. Canny, A computational approach to edge detection, IEEE Trans. PAMI (6) (1986) 679–698.
[6] W. Shen, X. Wang, Y. Wang, X. Bai, Z. Zhang, DeepContour: a deep convolutional feature learned by positive-sharing loss for contour detection, CVPR, 2015.
[7] S. Xie, Z. Tu, Holistically-nested edge detection, CVPR, 2015.
[8] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, J. Tangg, Richer convolutional features for edge detection, IEEE Trans. PAMI 41 (8) (2019) 1939–1946.
[9] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, X. Bai, Richer convolutional features for edge detection, CVPR, 2017.
[10] J. He, S. Zhang, M. Yang, Y. Shan, T. Huang, Bi-directional cascade network for perceptual edge detection, CVPR, 2019.
[11] X. Soria, E. Riba, A.D. Sappa, Dense extreme inception network: towards a robust CNN model for edge detection, WACV, 2020.
[12] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, ICLR, 2015.
[13] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CVPR, 2015.
[14] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, C. Fan, Sa-Unet: spatial attention U-net for retinal vessel segmentation, ICPR, 2020.
[15] D.R. Martin, C.C. Fowlkes, J. Malik, Learning to detect natural image boundaries using local brightness, color, and texture cues, IEEE Trans. PAMI 26 (5) (2004) 530–549.
[16] A. Pablo, M. Michael, F. Charless, M. Jitendra, Contour detection and hierarchical image segmentation, IEEE Trans. PAMI 33 (5) (2011) 898–916.
[17] P. Dollr, C.L. Zitnick, Fast edge detection using structured forests, IEEE Trans. PAMI 37 (8) (2015) 1558–1570.
[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, ECCV, 2018.

[19] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, CVPR, 2017.
[20] X. Ren, Multi-scale improves boundary detection in natural images, ECCV, 2008.
[21] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, CVPR, 2018.
[22] J. Hu, L. Shen, S. Albanie, G. Sun, A. Vedaldi, Gather-excite: exploiting feature context in convolutional neural networks, NIPS, 2018.
[23] Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, GCNet: non-local networks meet squeeze-excitation networks and beyond, ICCV, 2019.
[24] J. Park, S. Woo, J.-Y. Lee, I. SoKweon, BAM: bottleneck attention module, BMVC, 2018.
[25] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, CVPR, 2019.
[26] J.-B. Cordonnier, A. Loukas, M. Jaggi, On the relationship between self-attention and convolutional layers, ICLR, 2020.
[27] P. Shaw, J. Uszkoreit, A. Vaswani, Self-attention with relative position representations, 2018. arXiv preprint arXiv:1803.02155.
[28] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon1, CBAM: convolutional block attention module, ECCV, 2018.
[29] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, CVPR, 2017.
[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, NIPS, 2017.
[31] Y. Cai, Y. Wang, 2020. Ma-Unet: an improved version of Unet based on multi-scale and attention mechanism for medical image segmentation. arXiv preprint arXiv:2012.10952.
[32] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert, Attention U-net: learning where to look for the pancreas, MIDL, 2018.
[33] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from RGBD images, ECCV, 2012.