Bilateral attention network for semantic segmentation

Dongli Wang $^1$ | Nanjun Li $^1$ | Yan Zhou $^1$ | Jizhen Mu $^2$

1 School of Automation & Electronic Information, Xiangtan University, Xiangtan, China
2 Shanghai Aerospace Control Technology Institute, Shanghai, China

Correspondence
Yan Zhou, School of Automation, and Electronic Information, Xiangtan University, Xiangtan 411105, China.
Email: yanzhou@xtu.edu.cn

Abstract
Enhancing network feature representation capabilities and reducing the loss of image details have become the focus of semantic segmentation task. This work proposes the bilateral attention network for semantic segmentation. The authors embed two attention modules in the encoder and decoder structures. Specifically, high-level features of the encoder structure integrate all channel maps through dense channel relationships learned by the channel correlation coefficient attention module. The positively correlated channels promote each other, and the negatively correlated channels suppress each other. In the decoder structure, low-level features selectively emphasize the edge detail information in the feature map through the position attention module. The feature expression of semantic segmentation is improved by feature fusion of the two attention modules to obtain more accurate segmentation results. Finally, to verify the effectiveness of the model, the authors conduct experiments on the PASCAL VOC 2012 and Cityscapes scene analysis benchmark data sets and achieve a mean intersection-over-union of 74.92% and 66.63%, respectively.

1 INTRODUCTION

Semantic segmentation is the basis of computer vision tasks such as autonomous driving, medical image processing and image retrieval. Its purpose is to segment a scene into different image areas and assign each pixel in the scene a corresponding semantic label. With the great success of fully convolutional neural networks (FCN) [1] in this task, a series of methods based on FCN have been proposed. However, due to the structure of an FCN, along with the continuous downsampling of the convolution and pooling layers, contextual information will be missing. This will have a great negative impact on pixel-level classification. To solve this problem, the pyramid scene parsing network (PSPNet) [2] uses a pyramid pool module to capture multiscale contextual information. The DeepLab series [3–5] adopts atrous convolution. The global convolutional network (GCN) [6] expands the convolution kernel to obtain more contextual information. However, this method collects information from surrounding pixels and cannot generate dense contextual information. In recent years, attention modules have been successful in fields such as natural language processing [7–11], speech recognition [12, 13], image inpainting [14–16] and image recognition [17–19]. The self-attention layer of work [8, 11] maps a query and a set of key-value pairs to an output. It can look at other words in its input sequence through the current word to find clues that better encode this word. The work [13] uses a multilayer perceptron (MLP)-style attention with weight feedback to learn the relationship between any elements in the speech sequence and reduce the error rate of speech recognition. The multistage attention network [14] introduces a special multistage attention module with cascaded attention in the two layers of decoding to ensure good results with exquisite details after patch swapping processes. In addition, the approach considers structure consistency and detail fineness. It can effectively utilize the background information to accurately restore the mask areas. Several works [20–26] have introduced attention mechanisms to the field of image semantic segmentation to obtain contextual information. Among them, the most commonly used method is to establish a channel semantic dependency model. The squeeze-and-excitation network (SENet) [27] obtains the channel attention map through global average pooling and fully connected layers. The convolutional block attention module (CBAM) [28] combines global average pooling and global maximum pooling with a $1 \times 1$ convolution operation to obtain an attention map. The dual attention network (DANet) [23] applies non-local approaches [29] to image segmentation.
using a self-attention mechanism, which results in a more powerful pixel-level representation. Each channel map can be regarded as a class-specific response, and different semantic responses are related to each other. The magnitude of the correlation is used as a weight to aggregate various channel features and improve specific semantic feature representation. In this method, all weights are positive during weighted fusion. This means that the relationship between all responses is regarded as positive correlation, and different semantic responses promote each other. However, there is not only a positive correlation but also a negative correlation between the various channel responses. For example, as the semantic response to an indoor scene category increases, the response to an outdoor scene category should be suppressed rather than enhanced. If the weight used for weighted fusion could not only indicate the magnitude of the response correlation but could also represent the positive or negative correlation between channel responses, then better results would be obtained. Inspired by this idea, we propose the channel correlation coefficient attention module (C3AM) to simultaneously learn the positive and negative dependencies between channel maps, and we use the weighted sum of all channel maps to update each channel map.

Another problem of the FCN-based model is that although the high-level features of the encoding network have a good effect on the semantic classification, the image details will be lost to a certain extent when the original resolution is restored. SegNet [30], RefineNet [31] and other U-shaped encoder–decoder networks [22, 32] have boundary refinement modules that use low-level information and high-level features for feature fusion multiple times to restore image details. However, the integration of high-level and low-level features by layer and the use of boundary refinement modules increase the computational cost. To solve this problem, we designed a position attention module (PAM), which can help a network to better learn the spatial edge details of low-level features so that high-level and low-level fusion have a better effect. Only one fusion operation is needed to obtain good results, which reduces the calculation cost. Finally, based on the above method, a semantic segmentation framework, the bilateral attention network (BiANet), is proposed; the validity of the model is verified on the PASCAL VOC 2012 [33] and Cityscapes [34] data sets and the mean intersection-over-union (MIOU) reaches 74.92% and 66.63%, respectively.

In summary, the main contributions of this paper can be summarized as follows:

1) The C3AM is proposed to learn the correlations between channels and improve the segmentation results.
2) The PAM is proposed to highlight the details of edges in low-level features and improve the segmentation results by fusing low-level features with high-level features.
3) A new network, BiANet, is proposed for image semantic segmentation by combining the C3AM and PAM, and it achieves good results on the VOC and Cityscapes benchmark tests.

The organization of the rest of the paper is as follows. In the next section, we review recent methods for semantic segmentation, attentional mechanisms and encoder–decoder structures. In Section 3, we present our methods. We present experimental results and analysis in Section 4. Finally, we conclude the paper in Section 5.

2 RELATED

2.1 Semantic segmentation

The FCN [1] was the first implementation of a fully convolutional network for semantic segmentation and has become an indispensable model for semantic segmentation. The methods to improve semantic segmentation effects that currently exist are as follows: [3, 6] obtains contextual information by expanding the receptive field. PSPNet [2] and DeepLabV3 [4] introduce a pyramid pooling module and an atrous spatial pyramid pooling (ASPP) module to obtain multiscale contextual information. The encoder–decoder structure [22, 31, 35–38] integrates semantic features of different levels to obtain contexts at different scales. ParseNet [39] uses global pooling to obtain the global representation context information. The pointwise spatial attention network (PSANet) [40] captures pixel relationships through relative position information in convolutional layers and spatial dimensions. The object context network (OCNet) [26] uses a self-attention mechanism and ASPP to obtain context dependency. The context encoding network (EncNet) [21] introduces a channel attention mechanism to capture the global background. In addition, [3, 41, 42] use a conditional random field (CRF), Markov random field (MRF) and other graphical models to optimize the segmentation results. Refs. [43, 44] introduce adversarial learning into the task of semantic segmentation. The object-contextual representation network (OCRNet) [45] calculates the feature expression of a group of object regions and then propagates these object region feature representations to each pixel according to the similarity between the object region feature representation and the pixel feature representation.

2.2 Attention module

The attention module is used to simulate remote dependencies and has been widely used in many tasks, including natural language processing [7–10], image recognition [17, 18], image inpainting [14–16] and speech recognition [12]. At present, attention modules are also increasingly used in semantic segmentation. SENet [27] and CBAM [28] model the channel relationship in the attention mechanism through pooling operations to enhance the representation ability of the network. Ref. [29] proposes a non-local module for the task of video classification. A large attention map is generated by calculating the correlation matrix between each spatial point in the feature map, and then the contextual information is aggregated. OCNet [26] and DANet [23] use the non-local approach and propose an attention module to collect contextual information. Unlike the above types of attention, we also refer to the non-local approach; we calculate the positive and negative correlation matrix between

WANG ET AL.
each channel and use weighted fusion to improve feature distinguishability.

2.3 Encoder–decoder architecture

DeconvNet [46] uses deconvolution layers to recover full-resolution predictions gradually. SegNet [30] and Bayesian SegNet [47] use unpooling operations to obtain better performance. RefineNet [31], while deep feature aggregation (DFA) [37], and the pyramid attention network (PAN) [22] merge high-level and low-level features many times to improve performance. Ref. [48] proposes the data-dependent upsampling (DUpsampling) method instead of bilinear difference for upsampling. The bilateral segmentation network (BiSeNet) [32] designs boundary refinement to retain more feature space information. EfficientFCN [49] generates a codebook in the decoder and encodes the codeword to capture global context information. DeepLabV3+ [5] takes advantage of the encoder–decoder architecture and atrous convolution and has achieved good results. We embedded a PAM in the decoding structure of DeepLabV3+, which helps the network to focus on the critical edge information of low-dimensional features and further supplements the spatial information that high-dimensional features lack.

3 METHOD

The structure of the BiANet is shown in Figure 1. We employ a ResNet-101 [50] pretrained on ImageNet [51] with the dilated strategy as the backbone. Note that we remove the last fully connected layer and use atrous convolution in the Res-4 block, thereby reducing the size of the final feature map to 1/16 of the image [3]. We embed a C3AM module on top of ResNet to obtain positive and negative correlations between the global channels and then optimize the local response through the weighted fusion method so that the overall pixel classification can achieve better results. At the same time, low-level features are sent to the PAM module. The PAM is committed to highlighting edge details and other information in low-level features so that it can better compensate for the spatial information lost by high-level features. Finally, we fuse the output of the two attention modules and generate the final prediction map after simple bilinear upsampling by a factor of 4.

3.1 Channel correlation coefficient attention module

Channel maps of high-level features can be regarded as class-specific responses. We reviewed the existing channel attention modules, SE-attention [27] and CBAM-attention [28], which adopt the method of using a pooling operation to disperse the attention map. However, this strategy only gives different weights to parts and ignores the relationships among parts. These relationships are essential for scene segmentation. DANet [23] and OCNet [26] adopt non-local approaches to enable a single feature from any channel to perceive the features of all other channels. However, in this method, the relationship between channel responses is regarded as a positive correlation, and different semantic responses promote each other. In fact, there is not only a positive correlation but also a negative correlation between various responses.

Covariance is helpful in learning the positive or negative correlation between two responses. If the trends of the two responses are consistent, then the covariance between them is positive. Otherwise, the covariance between the two variables is negative. However, covariance can only qualitatively determine positive correlation and negative correlation but cannot quantitatively measure the degree of correlation between two variables. The Pearson correlation coefficient can be used to

FIGURE 1 Overview of the bilateral attention network (BiANet)
C3AM denotes channel correlation coefficient attention module. PAM denotes position attention module.
quantify the degree of correlation in the range of \([-1, 1]\). The larger the correlation coefficient is, the higher the correlation between the two. Inspired by this, we can learn the positive and negative correlations between channel responses by calculating the correlation coefficient matrix between channel responses and then improve the correlation representation of the network by weighted fusion. We refer to the module that learns these positive and negative correlations as the correlation coefficient attention module. The structure of the C3AM is shown in Figure 2. First, given the input feature \( A \in R^{C \times H \times W}\), we reshape it by channel to \( [A_1, A_2, ..., A_n] \in R^{C \times N}\), where \( A_i = [a_{i1}, a_{i2}, ..., a_{in}] \) is the set of the \( i \)th channel. \( a_{ij} \) denotes the response value of the \( j \)th position of the \( i \)th channel of the input feature. The relationship between any two channel responses can be obtained by the following formula:

\[
\mu_i = \frac{1}{n} \sum_{j=1}^{n} a_{ij}, \quad (1)
\]

\[
\text{cov}(A_i, A_j) = E((A_i - \mu_i)(A_j - \mu_j)^T), \quad (2)
\]

\[
v_{ij} = \text{cor}(A_i, A_j) = \frac{\text{cov}(A_i, A_j)}{\text{std}(A_i) \text{std}(A_j)}, \quad (3)
\]

where \( E(\cdot) \) denotes the expectation, \( \mu_i \) and \( \text{std}(A_i) \) denote the expectation and standard deviation of \( A_i \). The activation function commonly used in the attention model is the sigmoid, and the output value of the activation function is in \((0, 1)\), which contains only positive values. We want to learn both positive and negative correlations between channels at the same time and keep the value of the obtained relationship coefficient matrix in the range of \([-1, 1]\). Therefore, we do not use the activation function. \( v_{ij} \) is the impact of the \( i \)th channel response on the \( j \)th channel response. When it is positive, the \( i \)th channel response is positively correlated with the \( j \)th channel response. When it is negative, the channel response is negatively correlated. The larger the value is, the greater the correlation.

Finally, we perform matrix multiplication between \( V \) and \( A_i \) reshape the result to \( R^{C \times H \times W} \), and perform pixel-level addition with \( A \) to obtain the final output \( E \in R^{C \times H \times W} \).

\[
e_{ij} = \alpha \sum_{j=1}^{c} (v_{ij} a_{ij}) + a_{ij}, \quad (4)
\]

where \( \alpha \) is a learnable scale parameter starting from 0, \( e_{ij} \) denotes the response value of the \( j \)th position of the \( i \)th channel of the output feature of C3AM. After this operation, the final feature of each channel is the weighted sum of its original feature and all other channel features.

We use the Pearson correlation coefficient between channels as weights for weighted fusion. For a channel semantic response, the positively correlated response enhances its response, and the negatively correlated response inhibits the semantic response. The degree of this promotion or inhibition is determined by the magnitude of its correlation coefficient. The channel attention module can clearly learn the interdependence between channel maps. Our channel attention model makes full use of global channel information to optimize the response of each channel and enhances the intraclass consistency and interclass correlation of responses, thereby helping to obtain a better pixel-level prediction feature representation.

We can also calculate the spatial correlation coefficient matrix in this way to collect the spatial context information of high-level features. However, the computational cost of the \((H \times W) \times (H \times W)\) attention map is too large. Therefore, we did not use two attention modules for high-level features at the same time as in DANet.

3.2 | Position attention module

High-level features have rich semantic information. We can directly restore the output of the C3AM to the input size through bilinear upsampling by a factor of 16. However, this method does not restore the edge details of the object very well. Low-level features have rich spatial information and retain more details. A fusion of low-level and high-level features can better address the edges and details of the image. DeepLabV3+ [5] proposes a simple but effective decoding structure. As shown in Figure 1, we refer to the decoder structure of DeepLabV3+ and embed a PAM before fusing the low-level and high-level features.

The structure of the PAM is shown in Figure 3. CBAM [28] has shown that applying the average pooling and maximum pooling operations along the channel axis is effective in highlighting the information area. First, we perform the average pooling operation and maximum pooling along the channel axis. This operation yields two maps: \( F_{\text{avg}} \in R^{C \times H \times W} \) and \( F_{\text{max}} \in R^{C \times H \times W} \). Unlike CBAM, which directly concatenates the two and uses the convolution operation to obtain the position

FIGURE 2 The details of the channel correlation coefficient attention module
FIGURE 3  The details of the position attention module

attention map, we perform depthwise convolution on them before the convolution operation. When the features are obtained by directly concatenating the two pooling operations and then performing the convolution operation, due to the weight sharing of the convolution, the proportion of the two is the same. The features of the average pooling feature and the maximum pooling feature in the task of highlighting the information area do not necessarily have the same degree of contribution to the task of compensating for the edge details. Analogous to the classification task, the classifier will learn different weights for each attribute of the sample to find the best separating hyperplane. Depth convolution can be regarded as giving different weights to each pooling operation so that the network can better learn edge details. In short, the position attention map calculation process is as follows:

$$F_j = \sigma \left( f^{1x1} \left( d f^{1x1} \left( \left[F_{avg}, F_{max} \right] \right) \right) \right)$$

where $\sigma(\cdot)$ denotes the activation function, $d f^{1x1}$ denotes $1 \times 1$ depthwise convolution and $f^{1x1}$ denotes $1 \times 1$ convolution. $\omega_0$ and $\omega_1$ represent the weights of the two pooling operations. The attention map $F_j$ that we obtain is the weighted fusion of the information extracted by the two pooling operations. Finally, we take the Hadamard product of the obtained attention map and the original feature to obtain the desired compensation feature $G$.

4 | EXPERIMENTS

In this section, we first introduce the data set used for the evaluation and the specific details of the experiment. Then, we describe the results of a series of experiments carried out by using BiANet on the PASCAL VOC 2012 [33] and Cityscapes [34] data sets.

4.1 | Data sets and implementation details

PASCAL VOC 2012. This data set contains characters, vehicles, animals and indoor scenes—a total of 20 categories of foreground objects and a background class. There are 10,582 images for training in total and 1449 and 1456 for verification and testing, respectively.

Cityscapes. This data set contains street scenes from 50 different cities, with a total of 19 foregrounds and a background class. There are 2979 images for training, 500 images for verification and 1525 images for testing. Note that we do not use rough data in the experiment.

Our network is based on the PyTorch1.0 framework and trained on two GTX-1080Ti GPUs with 11 GB of memory. All experimental results in this section are based on this platform. In the network training process, we use the poly-learning strategy, and the initial learning rate is multiplied by $(1 - \text{iter} / \text{max_iter})^{0.9}$ each time. For PASCAL VOC 2012, we set the initial learning rate to 0.007 and batch size to 8. For Cityscapes, the initial learning rate, crop size and batch size are set to 0.01, 768 and 4, respectively. We train for 60 epochs on each data set. At the same time, we use small-batch stochastic gradient descent (SGD) as the optimization program, with the momentum set to 0.9 and the weight attenuation coefficient set to 0.0001. To prevent overfitting, we add a dropout layer to the output of the attention network.
We also compared our channel attention module with existing methods. We added the attention module proposed in [23, 27, 28] at the top of the dilated FCN-16 and experimented with the same settings. The results are shown in Table 2. The C3AM achieved better results than existing methods, even better than DANet [23] using two parallel modules at the same time.

When tanh is used as the activation function of the attention map, the output is also in the range of $[-1, 1]$. We also use tanh as the activation function and conduct experiments. The results show that compared with the weights learned after using the activation function, directly using the Pearson correlation coefficient can indeed better represent the channel relationship.

The effect of C3AM can be seen in Figure 4. Through our channel attention module, some misclassified categories are now correctly classified. Our model enhances the consistency of semantic responses and obtains better pixel-level prediction feature representation.

### TABLE 1
Ablation study of C3AM on the PASCAL VOC 2012 validation set

| Method            | BaseNet | MIOU % |
|-------------------|---------|--------|
| Dilated FCN-16    | Res50   | 69.92  |
| Dilated FCN-16+C3AM | Res50  | 71.37  |
| Dilated FCN-8     | Res50   | 70.12  |
| Dilated FCN-8+C3AM | Res50  | 72.06  |
| Dilated FCN-16    | Res101  | 71.51  |
| Dilated FCN-16+C3AM | Res101 | 73.66  |
| Dilated FCN-8     | Res101  | 72.10  |
| Dilated FCN-8+C3AM | Res101 | 73.96  |

### TABLE 2
Result on the PASCAL VOC 2012 validation set with different attention modules

| Method                        | BaseNet | MIOU % |
|-------------------------------|---------|--------|
| Dilated FCN-16                | Res101  | 71.51  |
| Dilated FCN-16+SE-attention   | Res101  | 72.29  |
| Dilated FCN-16+CBAMC          | Res101  | 73.09  |
| Dilated FCN-16+CBAMS          | Res101  | 73.10  |
| Dilated FCN-16+CBAM [28]      | Res101  | 73.13  |
| Dilated FCN-16+Bise-attention [32] | Res101 | 73.22  |
| Dilated FCN-16+DAP-attention  | Res101  | 72.64  |
| Dilated FCN-16+DAC-attention  | Res101  | 72.85  |
| Dilated FCN-16+DA-attention [23] | Res101 | 72.96  |
| Dilated FCN-16+C3AM(tanh)     | Res101  | 73.12  |
| Dilated FCN-16+C3AM           | Res101  | 73.66  |

SE-attention denotes the attention module published in SENet [27]. CBAMC denotes the channel attention module published in CBAM, CBAMS denotes the spatial attention module published in CBAM, CBAM denotes the entire attention module published in CBAM. Bise-attention denotes the attention module published in BiseNet. DAP-attention denotes the position attention module published in DANet. DAC-attention denotes the channel attention module published in DANet. DA-attention denotes the entire attention module published in DANet.

### 4.2.2 Ablation study for the PAM

We embedded a PAM in the decoding network to make the network learn more low-level spatial information. To verify the performance of the PAM module, we used the settings in Table 3 for testing.

As shown in Table 3, when the low-level features are directly integrated with the high-level features without embedding the
TABLE 3  Ablation study of the PAM on the PASCAL VOC 2012 validation set

| Method                                      | BaseNet | MIOU % |
|---------------------------------------------|---------|--------|
| BiANet (without PAM)                        | Res101  | 73.87  |
| BiANet (PAM -CBAMS) [28]                    | Res101  | 74.26  |
| BiANet (PAM -only GAP)                      | Res101  | 74.38  |
| BiANet (PAM -only GMP)                      | Res101  | 74.52  |
| BiANet (PAM -without depthwise convolution) | Res101  | 74.29  |
| BiANet                                       | Res101  | 74.92  |

GAP denotes global average pooling, GMP denotes global maximum pooling.

Attention mechanism, the MIOU of the model has only a 0.21% improvement. At the same time, we observe that the results obtained by using global average pooling alone to extract features are different from those obtained using global maximum pooling alone. If we directly concatenate the features obtained by the two and then perform the convolution operation to obtain the attention map, the obtained effect cannot even improve upon that using only one of them. This indicates that the characteristics of the two pooling operations in the task should be different. We use deep convolution to obtain the two different weights, and the resulting MIOU is 74.92%, which yields a 1.26% improvement in the network. In addition, our proposed module achieves better results than [23, 27, 28].

At the same time, Figure 5 shows that through our PAM, the edge detail information of the picture is improved to a certain extent.

TABLE 4  Comparison of the experimental results for the PASCAL VOC 2012 validation set

| Method                                      | BaseNet | MIOU % |
|---------------------------------------------|---------|--------|
| Dilated FCN                                 | Res101  | 71.51  |
| PSPNet [2]                                  | Res101  | 72.72  |
| DeepLabV3 [4]                               | Res101  | 73.45  |
| DeepLabV3+ [5]                              | Res101  | 74.26  |
| DANet [23]                                  | Res101  | 72.96  |
| OCNet (Bised-OC) [26]                       | Res101  | 72.64  |
| OCRNet [45]                                 | Res101  | 74.34  |
| EfficientFCN [49]                           | Res101  | 73.78  |
| BiANet                                      | Res101  | 74.92  |

4.2.3  Comparison with existing methods

Finally, we compare our BiANet with some existing methods on the PASCAL VOC 2012 validation set, and the results are shown in Table 4. In addition, we also show the segmentation prediction results in order to make an intuitive comparison, as shown in Figure 6.

4.3  Experimental results on the Cityscapes data set

We conducted experiments on the Cityscapes data set to further evaluate the effectiveness of our method. The quantitative results for the Cityscapes test set are shown in Table 5.
Our attention module significantly improves performance, with C3AM contributing a 17.32% improvement to the benchmark network. The PAM module produces a 1.24% improvement to the overall network.

At the same time, we compared BiANet with existing methods on this data set. The results show that our model achieves 66.63% for the MIOU. The two attention modules we introduced capture the global channel dependence and selectively emphasize low-level spatial details. The proposed method can achieve better performance. Similarly, we show the segmentation prediction results for an intuitive comparison, as shown in Figure 7.

5 | CONCLUSION

This paper proposed a BiANet for scene segmentation. Specifically, we introduced channel attention to obtain the global
dependencies of the channels of high-level features. At the same time, we introduced positional attention to focus on the spatial information of low-level features. The experiments showed that the two attention modules can effectively improve the accuracy of the segmentation results. Our proposed network achieved excellent performance on two scene segmentation data sets, namely, PASCAL VOC 2012 and Cityscapes.

ACKNOWLEDGEMENTS
This work was supported by the National Natural Science Foundation of China (61773330), the Project for Hunan National Applied Mathematics Center (20202YT003), the Project of Shanghai Municipal Science and Technology Commission (19511120900) and the Research Project of Department of Education of Hunan Province (19C1740).

CONFLICT OF INTEREST
No

PERMISSION TO REPRODUCE MATERIALS FROM OTHER SOURCES
None.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID
Yan Zhou https://orcid.org/0000-0002-2372-4947

REFERENCES
1. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440 (2015)
2. Zhao, H. et al.: Pyramid scene parsing network. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6230–6239 (2017)
3. Chen, L.C., et al.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40(4), 834–848 (2018)
4. Chen, L.C. et al.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:170605587 (2017)
5. Chen, L.C. et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 833–851 (2018)
6. Peng, C. et al.: Large kernel matters—Improve semantic segmentation by global convolutional network. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1743–1751 (2017)
7. Lin, Z. et al.: A structured self-attentive sentence embedding. In: Proc. Int. Conf. Learn. Representations (LCLR) (2017)
8. Vaswani, A. et al.: Attention is all you need. In: Adv. Neural Information Processing Systems, pp. 5998–6008 (2017)
9. Shen, T. et al.: DSAN: Directional self-attention network for RNN/CNN-free language understanding. In: AAAI Conf. Artif. Intell., pp. 5446–5455 (2018)
10. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proc. Int. Conf. Learn. Representations (LCLR) (2015)
11. Zadeh, A.H., Moshtorvos, A.: GOBO: Quantizing attention-based NLP models for low latency and energy efficient inference. In: 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO) (2020)
12. Chorowski, J. et al.: Attention-based models for speech recognition. In: Adv. Neural Information Processing Systems, pp. 577–585 (2015)
13. Rossenbach, N. et al.: Generating synthetic audio data for attention-based speech recognition systems. In: ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7069–7073 (2020)
14. Wang, N. et al.: Multistage attention network for image inpainting. Pattern Recognit. 106, 107448 (2020)
15. Liu, H. et al.: Coherent semantic attention for image inpainting. In: Proc. IEEE Int. Conf. Comput. Vision. Version (ICCV), pp. 4170–4179 (2019)
16. Yu, J. et al.: Generative image inpainting with contextual attention. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 5505–5514 (2018)
17. Mnih, V. et al.: Recurrent models of visual attention. In: Adv. Neural Information Processing Systems, pp. 2204–2212 (2014)
18. Gao, C., Zou, Y., Huang, J.B.: iCAN: Instance-centric attention network for human-object interaction detection. In: Proc. British Machine Vision Conference (BMVC), p. 41 (2018)
19. Cai, R., et al.: Skeleton-based attention-aware spatial-temporal model for action detection and recognition. IET Comput. Vision 14(5), 177–184 (2020)
20. Huang, Q. et al.: Semantic segmentation with reverse attention. In: Proc. British Mach. Vis. Conf. (BMVC) (2017)
21. Zhang, H. et al.: Context encoding for semantic segmentation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 7151–7160 (2018)
22. Li, H. et al.: Pyramid attention network for semantic segmentation. In: Proc. British Mach. Vis. Conf. (BMVC), p. 285 (2018)
23. Fu, J. et al.: Dual attention network for scene segmentation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 3146–3154 (2019)
24. Zhou, E., et al.: Flow driven attention network for video salient object detection. IET Image Proc. 14(6), 997–1004 (2020)
25. Zhang, J. et al.: Flow driven attention network for video salient object detection. IET Image Proc. 14(6), 997–1004 (2020)
26. Yuan, Y., Wang, J.: OCNet: Object context network for scene parsing. arXiv preprint arXiv:180900916 (2018)
27. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018)
28. Woo, S. et al.: CBAM: Convolutional block attention module. In: Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 3–19 (2018)
29. Wang, X. et al.: Non-local neural networks. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 7794–7803 (2018)
30. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39(2), 303–338 (2017)
31. Lin, G. et al.: RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 5168–5177 (2017)
32. Yu, C. et al.: BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 7151–7160 (2018)
33. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39(2), 303–338 (2017)
34. Cordts, M. et al.: The Cityscapes dataset. In: CVPR Workshop on the Future of Datasets in Vision (2015)
35. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241 (2015)
36. Yu, C. et al.: Learning a discriminative feature network for semantic segmentation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 1857–1866 (2018)
APPENDIX

The proof that the value of Pearson’s correlation coefficient is in the range of \([-1, 1]\) is as follows:

\[
E(x + y)^2 = E(x^2) + 2E(xy) + E(y^2),
\]

\[
\Delta = 4[E(xy)]^2 - 4E(x^2)E(y^2) \leq 0,
\]

\[
[E(xy)]^2 \leq E(x^2)E(y^2).
\]

Therefore, we obtain:

\[
\left[ E((A_i - \mu_i)(A_j - \mu_j)^T) \right]^2 \leq E((A_i - \mu_i)(A_j - \mu_j)^2),
\]

\[
-\text{std}(A_i)\text{std}(A_j) \leq \text{cov}(A_i, A_j) \leq \text{std}(A_i)\text{std}(A_j)
\]

\[-1 \leq \text{cor}(A_i, A_j) \leq 1.\]