An Introductory Survey on Attention Mechanisms in Computer Vision Problems

Jiahao Sun, Jie Jiang, Yang Liu,
College of Systems Engineering, National University of Defense Technology
Changsha, China
Email: {sunjiahao, jiejiang, yangliu}@nudt.edu.cn

2020 6th International Conference on Big Data and Information Analytics (BigDIA) | 978-1-6654-2232-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/BIGDIA51454.2020.00054

***Abstract*—Attention mechanism is a simple method derived from the research of human vision. With the development of artificial intelligence, it is gradually applied to the field of deep learning, especially computer vision. The weight of the pixels is redistributed in the form of a mask to compose attention mechanism. Various tasks in the field of computer vision, including image classification, object detection, image semantic segmentation, and fine-grained classification, have used attention mechanism widely and significantly improved. This paper presents an introductory summary of the mechanisms in computer vision problems through recent research. Provide readers with the basics of this widely used method, explore its manifestations under different tasks, evaluate the performance of attention mechanisms, and look forward to its development trends.**

***Index Terms*—Computer Vision; Attention Mechanisms**

I. Introduction

From the study of human visual attention mechanism in cognitive science, people in information processing bottleneck, selective attention to some information, while ignoring other visible information [1]. This mechanism is usually referred to attention mechanism [2] are no strict mathematical definition, form and mechanism of attention in the field of computer vision is usually through mask way to change the weight of local information.

After convolution neural network is put forward, the image classification and target detection has made huge progress [3] by convolution neural network. The images are first input into the neural network model to complete the extraction of their deep features, and then the results are output through the full connection layer to achieve the purpose of classification. Although the convolutional neural network has made a great breakthrough in image classification, it still has many shortcomings, mainly include: Firstly, the convolutional neural network will bring huge computational overhead in processing large images, which requires high hardware requirements; Secondly, in the process of convolution, the influence of global information will be ignored because convolution is the processing of local information. To solve the first problem, Mnih et al. [4] proposed Recurrent Models of Visual Attention, which is a new Recurrent neural network model based on the convolutional neural network. It processes input sequentially, processing different locations of an image one at a time, and gradually combining information from those locations to establish dynamic connections between scenes or environments. To solve the second problem, Wang et al. [5] proposed a module of non-local Neural Networks, which is a self-attention mechanism module and can compensate for the effect of ignoring the whole on a pixel in the convolution process.

So far, the adjustment space of the network structure of deep learning has been gradually reduced, and more scholars have introduced the attention mechanism into deep learning networks and achieved better results [6]. As for the attention mechanism, it is usually divided into soft attention mechanism [7] and hard attention mechanism [1]. The main difference between the two attention mechanisms is that the hard attention mechanisms are masked by 0 and 1, while the soft attention mechanisms are masked by 0 to 1. It is difficult to classify the current attention mechanisms strictly according to the mechanisms of soft attention or hard attention. In addition to target detection and key location extraction, most of the attention mechanisms are soft attention mechanisms. The essence of the attention mechanism is to locate the information of interest and suppress the influence of useless information. The attention mechanism has achieved a good effect in the field of computer vision and has been widely used in many computer vision tasks [8]. At present, there is no strict classification standard for attention mechanism, and there are usually many variations for different tasks. An overview of the following sections of this article:

(1) The basic form of attention mechanism in image classification task.

(2) A variety of attention mechanisms for different tasks.

(3) Analyze and evaluate the effectiveness of the attention mechanism.

(4) Look forward to the development of attention mechanism.

II. Basic form

After the emergence of neural network, different network architectures such as convolutional neural network(CNN) and recurrent neural network(RNN) emerged because of the different tasks they handled. On the basis of these architectures, the attention mechanism of image classification task is mainly divided into the attention mechanism based on CNN and the attention mechanism based on RNN.

In order to achieve this goal, the hard attention mechanism is generally used in the target detection area, which is to set

978-1-6654-2232-1/20/$31.00 ©2020 IEEE
DOI 10.1109/BigDIA51454.2020.00054

the weight of the area of concern to 1 and the other areas to 0.

Liu et al. [9] proposed SSD, which generated multiple default boxes on different feature graphs. Assuming that the size of the feature graph was $m \times n$, k default boxes were generated on a feature graph, and one default box corresponds to C categories and 4 border information, thus generating features with dimensions $m \times n \times k \times (4+c)$. The default box is a representation of the attention mechanism.

Then Ren et al. [10] proposed a new candidate box selection mechanism, RPN, which introduced the probability of Anchor relative to SSD. The default box on SSD is for each pixel, but in general, such boxes should appear around the foreground. RPN predicts whether this point is foreground on the basis of anchor, and produces k default boxes around it when it is foreground, thus reducing the total number of default boxes in a feature map. But in general, these are the general patterns of target detection and most of them don't think of it as an attention mechanism.

With the improvement of image recognition ability brought by the convolutional neural network, some deficiencies of the convolutional neural network are also exposed.

For example, in the classification of birds, the problem of fine-grained classification is extremely challenging because birds differ only in subtle and local differences. Therefore, Xiao et al. [11] proposed two-level Attention Models, which include object-level Attention Model and part-level Attention Model. The Object-Level Attention model uses the CNN trained on the target data set as a filter to screen out the Patches that contain the target of the parent class, and then re-input these Patches to DomainNet for training. On the other hand, the partial-level attention model only keeps three most influential patches for the selected Patches through clustering with AlexNet [3]. The application of attention mechanism in the problem of fine-grained classification is a process of selecting the focus area, and this mode is also the basic mode of attention mechanism in CNN.

For the RNN based attention mechanism, the understanding of the whole picture is accomplished by repeating over and over again. Mnih et al. [4] proposed a RAM module that, when a picture is entered, does not process a picture, but processes a piece of adaptive selection (Glimpse Window), processes a part of it in each cycle and sets up an incentive mechanism to integrate it when the classification is correct. The main contributions of this article are: Firstly, you can control the number of parameters and the amount of RAM performed independently of the size of the input image. Secondly, the model can ignore the irrelevant information in the image by placing its retina in the center of the relevant region. The attention mechanism based on RNN is a process from the part to the whole. The purpose of understanding the whole picture is achieved through the cognition of the part and the combination of these cognition.

In the field of computer vision, the expression form of attention mechanism is not uniform. This section introduces the basic form of attention mechanism for CNN-based fine-grained classification problem with representative significance and RNN-based image classification problem. This lays the foundation for a later introduction to variations in attention mechanisms that address different problems.

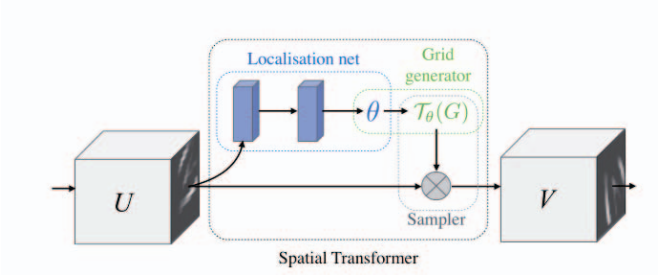

Fig. 1. STN network model [12]

III. Attention Mechanisms for Different Tasks

Attention mechanism has been widely used in different fields of computer vision. In the problem of picture classification, attention mechanism is used to extract key areas and recognize pictures through spatial invariance. In semantic segmentation, context connection is captured by adaptive attention mechanism to achieve better segmentation effect. Compared with the architecture of deep learning network, the attention mechanism is a lightweight module. Even in some architectures, the attention mechanism module exists as an embeddable functional module.

Through human cognitive intuition and a large number of experimental demonstration, the attention mechanism can solve the problem of missing important information in an image with too large size, the problem of information loss due to convolution, and the impact of global information loss on a pixel in the process of convolution. In this paper, we summarize the typical attention mechanisms, such as spatial invariant attention mechanism, self-attention mechanism, channel attention mechanism and other attention mechanisms.

A. Spatial invariance attention mechanism

When the classic CNN convolutional network is used for image classification, if the same object is shot from different angles, the recognition effect is particularly unsatisfactory. CNN lacks robustness for spatial invariance. An ideal image classification model would be one that still gives correct results after some transformation of the target.

Jaderberg et al. [12] proposed the STN model and effectively solved this problem by using the attention mechanism. It is also proved that STN can learn the invariance of pictures after translation, scaling, rotation and general deformation.

The STN network model, as shown in Figure 1, consists of three parts: Localisation Network, Grid Generator and Sampler.

Localisation Network is used to generate affine transform coefficients. Given an input $C \times H \times W$ image or characteristic graph U, learn a spatial transformation coefficient θ through

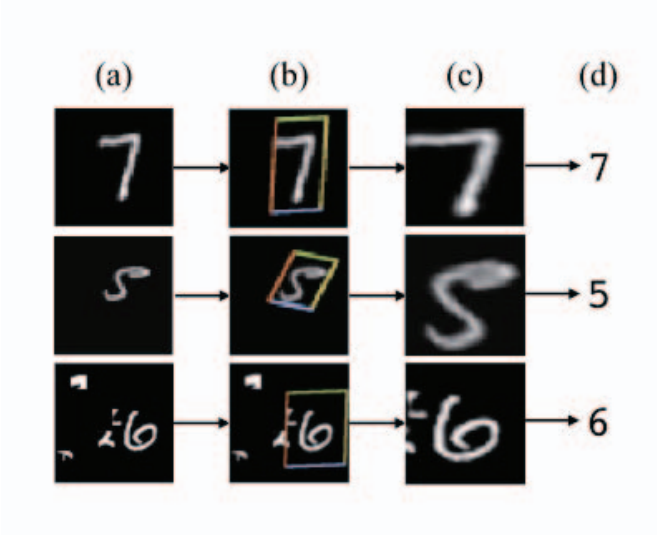

Fig. 2. Schematic Diagram of STN results [12]. Where (a) is the input image to the MNIST dataset, (b) is the area predicted by Localisation network to be transformed, (c) is the output of STN, and (d) is the classification result.

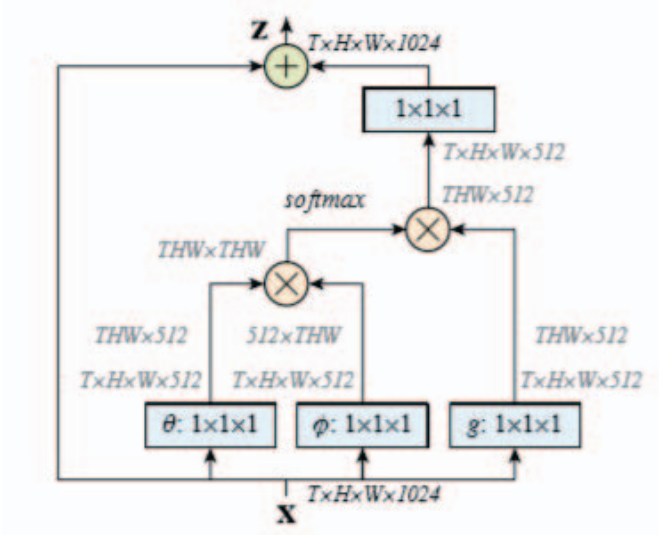

Fig. 3. Self-Attention Module Diagram [5]

convolution. The dimension of θ is determined according to the type of transformation.

The Grid Generator performs translation, rotation and other transformations on the input original image or feature graph based on the generated parameters. If the transformation is performed under 2D, the general formula of the transformation is as follows:

$$\begin{pmatrix} x_s^i \\ y_s^i \end{pmatrix} = T_\Theta\left(G_i\right) = A_\Theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_2 & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{1}$$

The sampler generates a new output image or feature image V based on the results obtained by the grid generator for the next operation.

STN is an independent module, which can be incorporated into the traditional CNN. Adding STN module to the picture classification task can change the picture, making it easier to identify and classify the changed picture, as shown in Figure 2.

But STN also has its limitations. Images with chaotic background images, complex scenes and large changes in appearance need to be modeled with different types of attention. And an attention template in STN can only modify one parameter.

In the application process of attention mechanism, the processing power of the network is generally enhanced by focusing on different locations or different representations of the same location. But human attention is complex, so there should be a tendency for different attention mechanisms to fuse, to fuse together to deal with problems. Based on this purpose, Wang et al [13] proposed Residual attention. It consists of two branches: the Mask branch and the Trunk branch. Each trunk branch has its own mask branch to learn about the areas or changes that different feature maps need to focus on. But blindly stacking attention modules will make the parameters of network structure too large and lead to problems such as overfitting. In order to solve this problem, they optimized their model through residual connection inspired by ResNet [14].

B. Autoattentional mechanism

In the process of convolution and pooling, because the convolution kernel is generally small in size, local processing is carried out, thus global information will be lost. In order to solve the problem of lack of global dependency in the process of convolution, the self-attention mechanism is proposed. When we look at an image, similar things in the image are related to each other. For example, buildings in different positions in an image are correlated. When we focus on buildings on the left side of the image, buildings on the right side of the image will also attract our attention. The mechanism of self-attention mainly studies the dependence between pixels with different positions in the space.

Vaswani et al. [15] proposed the self-attention mechanism in natural language processing. This model can be used to describe the dependence between words in a sentence. Subsequently, Wang et al. [5] applied the self-attention mechanism to the field of computer vision.Their attention model is used for video tasks. In an input feature map, the corresponding value of pixel X_i is assumed to be obtained by the weighted average of all X_j features.

$$y_i = \frac{1}{C\left(x\right)} \sum_{\forall j} f\left(x_i, x_j\right) g\left(x_j\right) \tag{2}$$

Formula 2 shows the calculation method of the ith output position. f is a function to calculate the relationship between i and all j, $g$is a transformation function to calculate the output characteristic representation of j position, and $C\left(x\right)$

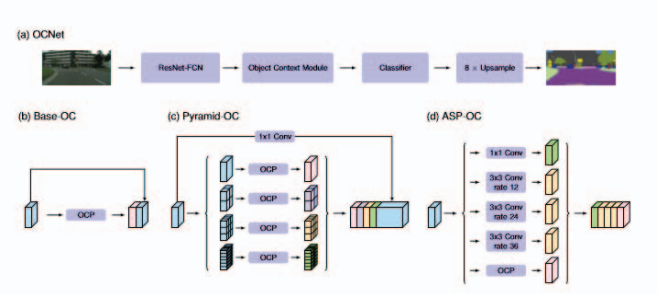

Fig. 4. OCNet attention module [16]

Fig. 5. DANet Visual Renderings [19]

is a normalization factor. In [5], four versions of F function, including Gaussian, Embedded Gaussian, Dot product and Concatenation, are proposed in order to calculate the correlation of different positions. These versions have been experimentally proved to be feasible, and the f function commonly used in subsequent studies is Dot Product, as shown in formula 3. The network structure of the attention module proposed in this paper is shown in Figure 3.

$$f\left(x_i, x_j\right) = \theta\left(x_i\right)^T \varphi(x_j) \tag{3}$$

X is the feature map of the neural network. After self-attention operation, it is added with the original X to get the final output.

In the contrast experiment with the residual module, it is proved that self-attention mechanism plays a significant role.The flexibility of spatial attention module, which can be easily inserted into neural network, is also discussed. When the network hierarchy is shallow it pays more attention to the correlation between larger objects, and when the network hierarchy is deep it pays more attention to the correlation between smaller objects.

Yuan et al. [16] defined the so-called Object context, that is, the set Object Context formed by the categories belonging to each position is defined by the feature similarity of different positions, that is, the similarity matrix in the process of self-attention. By multiplying the similarity matrix with the original feature, the Object Context can be applied to the feature graph.

As shown in Figure 4, base-oc is the basic self-attention module. Pyramid-oc and asp-OC combine self-attention with PSP module [17] and ASPP module [18] respectively. When extracting Object context, use the Pooling operation with different multiplicity or the void convolution with different ratios to obtain multi-scale features, and use the context information to segment the original image to the maximum extent.

Fu et al. [19] improved the self-attention mechanism and used the self-attention mechanism module in the task of image semantic segmentation. When FCN [20] restored the scale of the feature map to 1/8 of the original figure, spatial attention module and channel attention module based on self-attention mechanism were introduced to capture the global dependency and the dependency between channels. As a result, Cityscapes data sets [21] improved IOU scores by about 5% compared to networks without attention modules. In addition, the effect of self-attention mechanism is presented in the form of pictures through visualization technology, as shown in Figure 5. In the first row, in the first figure, a building is marked in red dot 1, and the obvious finding is that in the second column, the attention mechanism emphasizes the building area. The experimental results also show that the spatial attention mechanism can capture similar semantic information and ignore the distance between them.

The self-attention mechanism is intuitively and theoretically better for tasks that are globally dependent, so it performs well in tasks such as street view segmentation.

C. Channel attention mechanism

In general, each image is initially represented by (R,G,B) channels. After different convolution kernels, new signals will be generated for each channel. For example,32 new channel features (H,W,32) will be generated for each channel of image features by using 32 kernel convolution, where H and W respectively represent the height and width of features. The goal is to compensate for the loss of information by increasing the number of channels.

For a three-channel RGB image, if the image we want to identify is a red series, then we will definitely pay more attention to the value of the R channel. Similarly, for the different channels after convolution kernel, different weights should be set for them to focus on the channels more relevant to the key information. And this weight value can be obtained through deep learning. The greater the weight value, the more important the channel. In the process of using deep neural network to process images, depth features of images are extracted through continuous convolution, and these features are stored in the channel obtained by convolution.

Therefore, Hu et al. [22] proposed the weight value of SENet to find the channel. In other attention mechanisms, the idea of letting the network learn weights on its own is used, and the same idea applies to the channel attention mechanism. SENet learns to automatically capture the importance of each

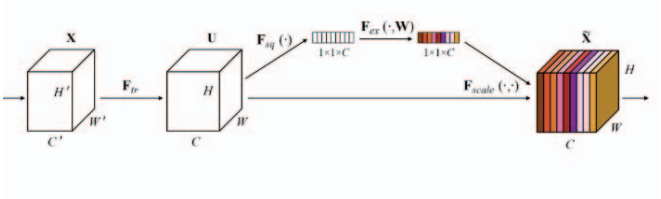

Fig. 6. SENet network Structure [22]

feature channel, and then uses that importance to enhance useful channel features and suppress channel features that are not useful for the current task. The SENet network structure is shown in Figure 6.

A Squeeze-Excitation module was introduced for SENet. The goal is to get the channel weights. Firstly, it performs the Squeeze operation, and compresses the spatial dimension to obtain a number with global receptive field. C real Numbers can be obtained by compressing C channels, and SENet selects global average pooling operation for Squeeze. Secondly, it performs the Excitation operation. After C real Numbers are obtained, a full FC connection layer is added to change the original channel into C/R. Then it is activated by ReLU function and restored to C channels through a full connection layer. The Sigmoid activation function is used to obtain the weight parameter W of each channel, which represents the difference between different feature channels.

SENet introduced a hospital-crawler-Branch for convolution.The purpose is to get the weight value of the channel.First, it performs the Squeeze operation to compress the spatial dimensions. After obtaining multiple feature graphs, it adopts each channel to convolve with a convolutional layer with the same size as the feature graph to obtain a real number with global receptive field.C real Numbers can be obtained by compressing C channels, and SENet selects global average pooling operation for Squeeze. Finally, it performs Scale operation. The obtained weight parameter W is weighted to the original feature by channel through multiplication to complete the original feature re-calibration on the channel dimension.

Woo et al. [23] proposed CBAM on the basis of SENet. There are two main changes: first, the Channel attention proposed in SENet is extended to Spatial attention. Spatial attention and Channel attention are connected through a serial branch, which enhances the original features.Second, in the pooling operation, considering that global average pooling will lose the prominent features of the image, the pooling layer in SENet is replaced by the combination of maximum pooling and global average pooling.

Zhang et al. [24] introduce a Context Encoding Module incorporating Semantic Encoding Loss (SE-loss). The Context Encoding Module use a fully connected layer on top of the Encoding Layer [25] to get channel weight. Then, the feature map is multiplied by the weights. In addition, the Semantic Encoding Loss allows the network to understand global semantic information with a very small additional computational cost.

Zhong et al. [26] proposed SANet. They argue that channel attention mechanisms should be combined with regional characteristics in semantic segmentation tasks. They use 1x1 convolution instead of the full connection layer after global average pooling. They also use up sampling instead of equivalent extension. The goal is to make the weights of each channel different according to the region.

D. The problem of fine-grained classification based on attention mechanism

In fine-grained classification problems, different categories can only be judged by local and subtle differences, and in order to achieve the classification of different categories, more attention to local and detail is needed. The attention mechanism can be used to highlight the difference between local and detail. However, attention mechanism also has different manifestations in fine-grained classification. For example, attention mechanism module is introduced on DNN to screen out more local photos conducive to classification and input them into the classification model. Or the attention mechanism module is introduced on the basis of RNN to achieve the purpose of classification from the local to the whole.

In the task of fine-grained classification, Xiao et al. [11] applied the Attention mechanism to this problem and proposed the two-level Attention Models. It can screen out three more conducive to classification of images, so as to classify. However, Mnih et al. [4] proposed a RAM module, which combines the attention mechanism and reinforcement learning [27]. Liu et al. [28] combined their approaches and used the idea of Markov decision (MDF) [29] to adaptively select multiple visual attention regions.

Fu et al. [30] introduced the idea of RPN into RNN network architecture and proposed RA-CNN. The method learns recursively to identify the regions to be noticed and the region-based feature representation in a mutually reinforcing manner. Each learning consists of a classification subnet and an attention suggestion subnet (APN). This network can gradually locate the most distinct areas from rough to fine, so as to enlarge different types of Loss and achieve better classification effect.

IV. CONCLUSION

Generally speaking, the attention mechanism mainly improves the effect by increasing the regional weight.

A brief summary of the attention mechanisms mentioned. For target detection tasks, the hard attention mechanism by cropping is a common method for almost all target detection algorithms. When the redundant information is removed, the classification effect can be significantly improved, which is obvious.

The purpose of attention module like STN, a space-oriented converter, is mainly to identify the same object from different angles and different perspectives. It is easier to identify the object through space transformation. The purpose of this method is mainly to remove the target differentiation performance caused by perspective and other reasons, so as to achieve the purpose of identification.

The self-attention mechanism is to enhance the global dependence of the image. In the process of convolution, both the convolution kernel and the pooling layer calculate the adjacent pixels, and the global information is lost.The self-attention mechanism is generated for the purpose of finding global dependencies, and its performance is better in the task of semantic segmentation.

Channel attention mechanism is because different feature information will be stored in different channels in the process of image processing, and the features of some channels will play a decisive role in image recognition. The relationship between channels should not be simple and equal, but prominent.This kind of attention mechanism is ideal for the task of classification and recognition due to shape or some typical features.

In addition, the attention mechanism based on RNN can make up for the memory information lost in the process of neural network convolution, and also achieve the purpose of local understanding of the whole.

REFERENCES

[1] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *Computer Science*, pp. 2048–2057, 2015.

[2] J. Anderson, "Cognitive psychology and its implications," 1980.

[3] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, no. 2, 2012.

[4] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[5] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," 2017.

[6] F. Wang and D. Tax, "Survey on the attention based rnn model and its applications in computer vision," *ArXiv*, vol. abs/1601.06823, 2016.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2015.

[8] J. Lee, R. A. Rossi, S. Kim, N. K. Ahmed, and E. Koh, "Attention models in graphs: A survey," *arXiv: Artificial Intelligence*, 2018.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," 2016.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[11] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 842–850.

[12] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[13] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," 2017.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.

[16] Y. Yuan and J. Wang, "Ocnet: Object context network for scene parsing," *ArXiv*, vol. abs/1809.00916, 2018.

[17] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239, 2017.

[18] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *ArXiv*, vol. abs/1706.05587, 2017.

[19] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3141–3149, 2019.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.

[21] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016.

[22] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 2011–2023, 2020.

[23] S. Woo, J. Park, J.-Y. Lee, and I. Kweon, "Convolutional block attention module," in *ECCV 2018*, 2018.

[24] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[25] H. Zhang, J. Xue, and K. Dana, "Deep ten: Texture encoding network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[26] Z. Zhong, Z. Q. Lin, R. Bidart, X. Hu, I. B. Daya, Z. Li, W. S. Zheng, J. Li, and A. Wong, "Squeeze-and-attention networks for semantic segmentation," 2019.

[27] R. Sutton and A. Barto, "Reinforcement learning: An introduction," *IEEE Transactions on Neural Networks*, vol. 16, pp. 285–286, 2005.

[28] X. Liu, T. Xia, J. Wang, and Y. Lin, "Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition," *ArXiv*, vol. abs/1603.06765, 2016.

[29] M. Puterman, "Markov decision processes: Discrete stochastic dynamic programming," in *Wiley Series in Probability and Statistics*, 1994.

[30] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4476–4484, 2017.