Saliency Prediction Based On Lightweight Attention Mechanism

Yu Wang¹,², Ming Zhu¹,²*
¹Electronics and Information Engineering, Anhui University, Hefei, Anhui, 230601 China.
²National Engineering Research Center for Agro-Ecological Big Data Analysis & Application, Anhui University, Hefei, AnHui, 230601
*Corresponding author’s e-mail: zhuming@ahu.edu.cn

Abstract: Saliency prediction refers to an algorithm that extracts salient regions from natural scenes. In the field of deep learning, receptive field limits the accuracy of pixel classification and thus affects the accuracy of saliency prediction. This study proposes a new saliency prediction method that uses the improved feature pyramid attention (FPA) to gain multi-scale context information to solve the above-mentioned problems. FPA's number of parameters and cost of calculation are also decreased. Experimental results show that this method can obtain more accurate results than the existing saliency prediction methods without increasing the calculation amount.

1. Introduction
The human visual system can quickly search and locate objects of interest, and these regions of interest that are automatically extracted by these people’s eyes are called saliency regions. In a complex natural environment, humans can understand the scene quickly by extracting saliency regions in the scene. Every aspect of information in the scene need not be observed before they can be understood. In the field of computer vision, the task of simulating human eye features to find saliency regions are called visual saliency prediction. Using this method of saliency prediction, key information can be easily extracted from large amounts of images, and the acquisition of useful information in a short time can be maximized to understand and simplify complex visual problems. Therefore, visual saliency prediction has important applications in many fields, such as video understanding [1], video compression [2], object recognition [3], and action recognition [4].

An input image is shown in Figure 1(a). The output of the saliency prediction is the ground truth fixation map shown in Figure. 1(b) or the ground truth density map shown in Figure. 1(c). The resulting graph shows that the saliency prediction needs to score all the pixels in the image according to human attention.
The result of the pixel-level classification directly affects the accuracy of the result of the saliency prediction. In the field of deep learning, the accuracy of pixel-level classification is difficult to improve due to the limitation of the size of the receptive field and the inaccuracy of spatial location. Thus, deep learning produces a saliency prediction result. Improving the accuracy of pixel-level classification results of neural networks is a major problem, which will improve the prediction results if solved. This study proposes the following innovations to solve the aforementioned problem:

(1) This study improves Feature Pyramid Attention structure and applies it to the saliency prediction to obtain accurate spatial location information and rich semantic information through multi-scale feature information. Accordingly, the influence of the receptive field limitation on the pixel-level classification can be solved.

(2) The attention mechanism is optimized through the idea of lightweight network to reduce the influence of large convolution kernel operation on model size and running speed.

2. Related work

2.1 Saliency prediction
In the 1980s, Treisman et al. [5] first proposed the principle of saliency prediction. A related study of saliency prediction begins. The first wave of research on saliency prediction began in 1998. Itti et al. [6] designed the first saliency prediction computational model by combining the low-level feature information, such as color, intensity, and direction, to produce the final saliency prediction map. The method of extracting and merging features becomes a general idea of the design of saliency computing models. Achanta et al. [7] and Liu et al. [8, 9] defined saliency prediction as a binary segmentation problem. Although the relationship between its definition and object detection is determined, the segmentation problem is relatively vague. As a result, the second wave of research begins. In recent years, the development of Convolutional Neural Networks (CNNs) has facilitated the progress of computer vision in all disciplines, and saliency prediction has ushered in the third wave of research. The application of CNNs in the saliency prediction is mainly divided into two models [10].

The first model is CNN-based model, which obtains the final saliency prediction maps directly from the high-level feature maps extracted by CNNs. This type of model uses small computational resources to obtain prediction results, but meticulously classifying each pixel is difficult and the high-level feature map lacks spatial information because of the limitation of the feature map receptive field. Wang et al. [11] alternately used top-down and bottom-up saliency predictive models through the characteristics of Long Short-Term Memory (LSTM) iterations to obtain accurate and saliency prediction from high-level feature maps. Marcella et al. [12] designed Saliency Attentive Model (SAM) to use LSTM and dilated convolutions in the saliency prediction model. The model uses LSTM iteration to obtain considerable spatial information, ensure the accuracy of spatial positioning, and use dilated convolution to obtain a large receptive fields for obtaining accurate prediction results.

The second model is Fully Convolutional Network (FCN)-based model. This model uses FCNs as base net. The low-level feature map is used to supplement the detailed information for ensuring accurate pixel-level classification and solving the pixel-level classification inaccuracy caused by the limitation in high-level feature map receptive field. However, the computational overhead and computation time of this model are greatly increased due to the increase in multi-level up sampling. Liu et al. [13] designed
PiCAnet with global attention, local attention, and U-net. Attention mechanism can help the network improve in gaining context information and spatial selection. Zhao et al. [14] designed the Pyramid Feature Attention Network (PFA), which uses Feature Pyramid Networks (FPN) with attention mechanism as the basic network. The method utilizes high-level semantic features and low-level spatial structure features to achieve accurate classification results.

Figure. 2 SAM structure. Dilated Convolutional Network is the base net. The structure is VGG16 or ResNet-50 with dilated convolution. AttentiveConvLSTM is Convolution Long Short-term Memory with activation function. Learned Priors refer to Gaussian priors.

2.2 Attention mechanism
A deep network depth corresponds to a wide width and a strong network classification performance. However, simply expanding the depth and width of the network may negatively affect the network and change the network structure needs. The overall network parameters are adjusted, and the workload is huge. The emergence of attention mechanism has introduced a new way to improve the performance of network classification. The attention mechanism can be easily embedded in most neural networks, and the network performance can be improved without changing the overall structure and calculation of the network. Li et al. [15] designed the Pyramid Attention Network (PAN) with attention mechanism and FPN. Feature Pyramid Attention (FPA) structure in PAN is shown in Figure.4 (a). This structure can acquire and integrate semantic information of different scales through the U-shape structure. The Global Pooling branch can obtain global information. Accordingly, the high-level feature map can directly generate improved pixel-level classification.

2.3 Lightweight network
The CNN model is robust, and its detection results are considerably better than those of the traditional algorithms in complex natural scenes. However, as the network continues to deepen and widen, the operating efficiency continuously reduces, and the equipment requirements and computing costs become increasingly high. These conditions limit the development of neural networks. The huge parameter quantity greatly affects the detection speed of the neural network, and this condition brings problems for practical applications. In recent years, an increasing interest has been given in building small and efficient neural networks. Many different methods can reduce the size of the model and increase its speed. Howard et al. designed the MobileNets [16] series to reduce the amount of calculation and keep the calculation efficiency. The ShuffleNets series [17] designed for lightweight neural networks specifically addresses the problem of mobile devices.

3. Improved algorithm
This study directly implements FPA for the high-level feature map and obtains multi-scale contextual and spatial positioning information directly from the high-level feature map. Convolutional Block Attention Module (CBAM) [18] is also embedded in FPA. The CBAM has two parts: channel and spatial attention. The fixed 7×7 convolution operation in the spatial attention is also changed for improved integration into the FPA structure. Channel attention mechanism can accelerate information flow and improve network anti-noise ability. Spatial attention mechanism can provide spatial information of different scales and can thus achieve accurate spatial positioning. CBAM adds the global maxpooling information that is lacking in the original FPA structure, optimizes the ability of the U-structure to acquire information, and utilizes feature information of different scales for improving space positioning ability.
The FPA module obtains feature maps of different scales by using large convolution kernels. However, the computational cost is still higher than that of the original network. The huge amount of calculation slows down the detection speed. Although this method obtains high accuracy, its excessive computational cost and model size are not conducive to its practical applications. The attention mechanism is a lightweight module. Thus, adding many parameter quantities does not match its concept. Accordingly, this study uses the idea of lightweight network to reduce the model parameters and calculation time as much as possible while ensuring the same accuracy.

For the original \( K \times K \times C \) convolution operation shown in Figure. 3, the parameter quantity is \( K \times K \times C \), and the improved parameter quantity is \( 2 \times 1 \times \frac{C}{2} \). The original is considered, except that the spatial separable convolution reduction parameter is replaced. This study merges depthwise separable convolution to reduce the size of the feature map and the number of parameters. As a result, the loss of precision and the number of parameters are decreased, and the amount of information in each channel is increased through Channel Shuffle. Ultimately, the overall loss of precision is decreased.

The application of the lightweight module and CBMA in the FPA structure is shown in Figure. 4(b). The convolution operation of the high-dimensional feature map using the large convolution kernel in the FPA structure is the main reason for the increased model parameter number and slow calculation speed. The convolution operation in the structure is lightweight, and the purpose of reducing the model parameters is achieved. The use of separable convolution greatly influences the accuracy. This effect can be minimized by adding Channel Shuffle at the joint of the branch structure.

The L-FPA module added to the SAM network is shown in Figure. 5. When the lightweight attention mechanism is used, the accuracy is slightly lower than that of the FPA-SAM model. However, its calculation speed and practical application are greatly increased or decreased.

The overall loss function consists of three parts, as calculated using Equation (1):

\[
L(\tilde{y}, y^{den}, y^{fix}) = \alpha L_1(\tilde{y}, y^{fix}) + \beta L_2(\tilde{y}, y^{den}) + \gamma L_3(\tilde{y}, y^{den}),
\]

where \( \alpha, \beta, \) and \( \gamma \) are the weights to balance the three loss settings; \( \tilde{y} \) is a network output saliency map; \( y^{den} \) represents the groundtruth density map; \( y^{fix} \) represents the groundtruth fixation map; \( L_1, L_2, \) and \( L_3 \) are the results of the evaluation parameters normalized scanpath saliency (NSS), linear correlation coefficient (CC), and Kullback–Leibler divergence (KL-DIV), respectively. The specific solutions of NSS and CC can be found in Section 4.1. KL-DIV measures the difference between two probability distributions in the same event space. The calculation is shown in Equation (2):

\[
L_3(\tilde{y}, y^{den}) = \sum_i y_{i}^{den} \log \left( \frac{y_{i}^{den}}{\tilde{y}_{i} + \epsilon} + \epsilon \right),
\]

where \( \epsilon \) is the regularization constant.
4. Experimental details and results

4.1 Evaluation parameters
We use the same four conventional evaluation parameters, namely, NSS, CC, area under the curve (AUC), and shuffle AUC (sAUC) as SAM, for predictions to evaluate model performance. The four evaluation parameters are detailed as follows.

(1) NSS [19]: NSS is a metric designed for saliency map evaluation, quantifies the saliency map values of gaze points, and normalizes them with saliency map variances. The calculation method is shown in Equation (3):

$$L_1(\hat{y}, y_{fix}) = \frac{1}{N} \sum_i (\hat{y}_i - \mu(y)) \cdot y_{fix},$$  \hspace{1cm} (3)

where $y_i$ is the $i$th pixel and $N$ is the binary truth map and the sum of all the pixels. $\mu$ is normalized to mean, and the unit standard deviation is zero.

(2) CC [20]: CC is a statistical method called Pearson correlation coefficient, which usually measures the correlation or dependence between two variables. CC can be used to evaluate the linear correlation between the predicted eye-point saliency map and the reference groundtruth map. The calculation method is shown in Equation (4):

$$L_2(\hat{y}, y_{den}) = \frac{\sigma(\hat{y}, y_{den})}{\sigma(y) \cdot \sigma(y_{den})},$$ \hspace{1cm} (4)

where $\sigma$ is the covariance among them.

(3) AUC/sAUC [21]: The area enclosed by the graph curve and the lower half is called AUC, and the ROC is false positive rate. For the horizontal axis, the true positive rate is the coordinate graph of the vertical axis. The threshold value of the predicted eye gaze is classified and plotted with different
4.2 Experimental details

This study tests the model performance on SALICON [22], MIT1003 [23], and CAT2000 [24] datasets. For all datasets, the batch size is set to 10, and the loss function weights \( \alpha, \beta, \) and \( \gamma \) are set to \(-1, -2,\) and 10. When the basic network is VGG16, the initial learning rate is \(10^{-4}\), and the learning rate is decreased by 10 times every two epoch. When the SALICON dataset is used for training, the image needs to be normalized to the size of 240×320. When MIT1003 is used for training, the image normalization size is still 240×320. However, the aspect ratio of the MIT1003 dataset is not 3:4 but 4:3. Thus, the image is preprocessed with 0 padding to keep the original aspect ratio of the image. When the CAT2000 dataset is used for training, the image normalization size is 180×320.

4.3 Experimental results and analysis

Table 1: Comparison of experimental results with SAM on SALICON, MIT1003, and CAT2000

| DATASETS | MODEL   | CC   | NSS  | AUC  | sAUC | SPEED |
|----------|---------|------|------|------|------|-------|
| SALICON  | SAM     | 0.830| 3.129| 0.883| 0.782| 0.32s |
|          | FPA     | 0.840| 3.293| 0.886| 0.786| 0.40s |
|          | Improved FPA | 0.840| 3.342| 0.887| 0.786| 0.42s |
|          | SDCov   | 0.825| 3.350| 0.889| 0.788| 0.34s |
|          | L-FPA   | 0.838| 3.350| 0.890| 0.788| 0.35s |
| MIT1003  | SAM     | 0.757| 2.852| 0.910| 0.613| 0.25s |
|          | FPA     | 0.773| 2.890| 0.915| 0.630| 0.30s |
|          | Improved FPA | 0.780| 2.948| 0.921| 0.636| 0.35s |
|          | SConv   | 0.760| 2.952| 0.925| 0.636| 0.26s |
|          | L-FPA   | 0.776| 2.950| 0.927| 0.636| 0.28s |
| CAT2000  | SAM     | 0.879| 2.347| 0.877| 0.530| 0.35s |
|          | FPA     | 0.882| 2.360| 0.875| 0.541| 0.39s |
|          | Improved FPA | 0.882| 2.416| 0.880| 0.541| 0.41s |
|          | SConv   | 0.875| 2.420| 0.882| 0.541| 0.36s |
|          | L-FPA   | 0.880| 2.422| 0.880| 0.538| 0.37s |

Table 1 shows the results of comparison of our model with the original SAM. As shown in the table, each evaluation parameter is improved after FPA is added. Therefore, high-order feature maps can obtain considerable context information through multi-scale feature fusion, and the detection results have high accuracy. The improved FPA module enhances the spatial positioning and noise immunity of the network. Thus, the accuracy of pixel classification gradually increases. Therefore, the improvement of NSS is significant. The module also helps the network accelerate information flow and quickly extract valid features. In the same experimental environment, SAM needs 10 epochs to iterate the final result, and each epoch requires 3200 s. The improved FPA module embedding requires only 6 epochs, and each epoch requires 4200 s. SAM only needs 400 MB, but the improved FPA model has more parameters and requires 1 GB of storage space. When L-FPA is used without channel shuffle, the prediction accuracy is slightly decreased. The reason is that the use of separable convolution indirectly leads to a small convolution kernel scale and less information in the branch. A reduced parameter amount also causes the function to be simulated. The ability to change is poor as well. CC decreases significantly, but the separable convolution extends the depth of the network to a certain extent. Thus, NSS and AUC have a certain degree of increase. When L-FPA is used with channel shuffle, CC has a significant increase in the amount of information in the branch. Meanwhile, NSS and AUC have a certain improvement. The overall results do not exceed those of the original improved FPA because the error caused by the reduction of the overall parameters is inevitable. In terms of speed and model size, the training time required for each era is only 3400 s while ensuring that the final result is 6 iterations. This model is only...
450 MB. Thus, the separation convolution can significantly reduce the computational cost and running time. The prediction accuracy is guaranteed at the same time. When channel shuffle is added, its accuracy improves, but the speed and model size do not change significantly.

| MODEL               | CC     | NSS    | AUC   | sAUC  |
|---------------------|--------|--------|-------|-------|
| Improved FPA +SAM   | 0.840  | 3.342  | 0.887 | 0.786 |
| L-FPA+SAM           | 0.838  | 3.350  | 0.890 | 0.788 |
| SAIGAN [25]         | 0.781  | 2.459  | 0.781 | 0.772 |
| DSCIRCN [26]        | 0.831  | 3.157  | 0.884 | 0.776 |
| DeepGazeII [27]     | 0.509  | 1.336  | 0.885 | 0.761 |
| DVA [28]            | 0.720  | 2.120  | 0.860 | 0.760 |

Table 2 compares the experimental results with the relevant recent findings under the SALICON dataset. As shown in the table, the improved network is superior related work.

5. Conclusions
We attempt to use the mainstream semantic segmentation method for saliency detection. When the FPA module is lightened and improved for saliency detection, considerable context and multi-scale feature information can be obtained without increasing the amount of calculation. Accurate pixel positioning can also be obtained. The loss of spatial information will not increase the computational burden. We use the CBAM module instead of the convolution operation in FPA to obtain considerable spatial information for guaranteeing pixel positioning, speeding up feature selection, and increasing anti-noise performance without affecting the classification. At the same time, the accuracy and convergence speed are improved, and the training time is decreased. Given that the anti-noise performance is improved, the obtained significant prediction graph is smooth and highly accurate. The performance of a few evaluation parameters has been improved, and good experimental results have been obtained.

Acknowledgments
This study is supported by the National Natural Science Foundation of China (No. 61501003) and National Key R&D Project (SQ2018YFC080102)

References
[1] D. Zhang, J. Han, L. Jiang, S. Ye, and X. Chang, Revealing event saliency in unconstrained video collection, [J]IEEE Trans. on Image Processing, vol. 26, no. 4, pp. 1746–1758, 2017.
[2] H. Hadizadeh and I. V. Bajic, Saliency-Aware Video Compression, [J] IEEE Transactions on Image Processing, vol. 23, no. 1, pp. 19–33, 2014.
[3] D. Gao and N. Vasconcelos, Discriminant saliency for visual recognition from cluttered scenes,[J]in Advances in Neural Information Processing Systems, 2004, pp. 481–488.
[4] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, X. Gao, Discriminative Multi-Instance Multi-Task Learning for 3D Action Recognition, [J] IEEE Trans. on Multimeddia, vol. 19, no. 3, pp. 519-529, 2017.
[5] Zheng Shou, Xudong Lin, Yannis Kalantidis, Laura Sevilla-Lara, Marcus Rohrbach, Shih-Fu Chang, Zhicheng Yan , DMC-Net: Generating Discriminative Motion Cues for Fast Compressed Video Action Recognition[C], in CVPR 2019
[6] Anne M. Treisman, Garry Gelade. A Feature-Integration Theory of Attention[C]. Cognitive Psychology, 1980, 12(1):pp.97-136.
[7] Itti L. A model of saliency-based visual attention for rapid scene analysis[C]. IEEE Trans. 1998, 20.
[8] R. Achanta, F. Estrada, P. Wils, and S. Susstrunk, Salient region detection and segmentation, [C] in Comp. Vis. Sys., 2008
[9] T. Liu, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum, Learning to detect a salient object,[C] in CVPR, 2007, pp. 1–8
[1] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, Learning to detect a salient object, [J] IEEE TPAMI, vol. 33, no. 2, pp. 353–367, 2011.
[11] Ali Borji, Ming-Ming Cheng, Qibin Hou et al. Salient object detection: A survey[J]. Computational Visual Media, 2019(2).
[12] W. Wang, J. Shen et al. An Iterative and Cooperative Top-down and Bottom-up Inference Network for Salient Object Detection[C]. in CVPR, 2019.
[13] Marcella C, Lorenzo B, Giuseppe S, et al. “Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model” [J]. IEEE Transactions on Image Processing, 2018. pp1-1.
[14] Liu et al. PiCANet: Learning Pixel-wise Contextual Attention for Saliency Detection[C], in CVPR, 2018.
[15] Zhao et al. Pyramid Feature Attention Network for Saliency Detection[C], CVPR, 2019.
[16] Li H, Xiong P, An J, et al. "Pyramid Attention Network for Semantic Segmentation" [C], in CVPR, 2018.
[17] Howard A G, Zhu M, Chen B, et al. "Searching for MobileNetV3, arXiv: https://arxiv.org/pdf/1905.02244.pdf
[18] Zhang X, Zhou X, Lin M, et al. (2019) ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design [C], in CVPR 2018.
[19] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module [C]. IN ECCV, 2018.
[20] R. J. Peters, A. Iyer, L. Itti, and C. Koch, Components of Bottom-Up Gaze Allocation in Natural Images, [J] Vision research, vol. 45, no. 18, pp. 2397–2416, 2005.
[21] Nahler G. linear correlation coefficient [J]. Encyclopedia of Systems Biology, 2009:1131-1131.
[22] Fawcett, T. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2006, 27(8): 861–874.
[23] Ming J, Huang S, Duan J, et al. SALICON: Saliency in Context[C], in CVPR IEEE, 2015.
[24] Judd T, Ehinger K, Frédo Durand, et al. Learning to predict where humans look[C] in 2009 IEEE 12th International Conference on Computer Vision. IEEE, 2009.
[25] Ali Borji, Laurent Itti, CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research[C], in CVPR, 2015.
[26] Pan J, Ferrer C C, Mcguinness K, et al. SalGAN: Visual Saliency Prediction with Generative Adversarial Networks [C]. in CVPR IEEE, 2019.
[27] N. Liu and J. Han, A Deep Spatial Contextual Long-term Recurrent Convolutional Network for Saliency Detection,[J] IEEE Transactions on Image Processing, vol. 27, no. 7, pp. 3264–3274, 2018.
[28] M. Kummerer, T. S. Wallis, L. A. Gatys and M. Bethge, Understanding low-and high-level contributions to fixation prediction,[C] in IEEE International Conference on Computer Vision, 2017.
[29] W. Wang and J. Shen. Deep visual attention prediction[J]. IEEE Transactions on Image Processing, Vol. 27, No. 5, pp 2368-2378, 2018.