PNGNET: A CAMOUFLAGE DETECTION NETWORK BASED ON PYRAMID POOLING MODULE

Xincheng GUO[1]*

In recent years, recent researchers like Fan et al. have developed methods based on deep learning such as PraNet and SINet, which contributed a lot to the field of camouflage object detection. However, there is a contradiction between the receptive field of feature extraction and the resolution of the feature map. To deal with the problem, we adopt the method of integrating the features of different receptive fields with the features of sub-regions. And this paper proposes a new deep neural network for camouflaged detection — PNGNet. By using the Pyramid Pooling Module, the semantic task of the image segmentation of the camouflage object is more accurate and the feature representation ability is enhanced. After times of experiments, the PNGNet outperforms most of the models from the perspective of three representative camouflage detection datasets with higher performance and robustness. The proposed network model takes into account both the receptive field of feature extraction and the resolution of the feature map and provides a new idea for camouflage detection. In future work, the results can also be applied in the medical field of segmenting pneumonia and polyps, image search, field rescue, and other fields.

Keywords: Concealed Object Detection, Pyramid Pooling Module, Deep Learning

1. Introduction

In nature, animals camouflage themselves by imitating the color, pattern, and brightness of their environment so as to reduce the risk of being detected by predators. The traditional object detection methods are mainly salient object detection (SOD) and general object detection (GOD). In contrast, camouflaged object detection (COD) based on background matching is more difficult in object recognition and segmentation. There are two main reasons for this: first, the camouflage can almost completely blend with the background, so it is hard to be distinguished from its surroundings; it lacks the strong contrast required by the segmentation method. Secondly, the shape, color, and size of all kinds of camouflage are different, so the training is difficult.

The early stage of camouflage object detection greatly relied on artificial feature extraction. However, when dealing with strong intra-class differences and the weak inter-class difference between camouflage objects and background regions, the representation ability of artificial feature extraction is quite limited. In

[1] School of Information Management, Wuhan University, Wuhan, China, e-mail: 2018302092009@whu.edu.cn, *corresponding author: Xincheng Guo

recent years, Fan et al. have developed PFANet, PraNet[1], SINet, and other methods based on depth learning to simulate the receptive field of the human visual system in the semantic task of image segmentation so as to enhance the ability of discriminative feature representations extraction[2]. However, as mentioned above, there is a contradiction between the receptive field of feature extraction and the resolution of the feature map. There are usually two methods to obtain a large receptive field. One is to adopt a large convolution kernel. However, this method will cost more computational resources. The second is using a large stride in the pooling, but this way will lose the resolution and greatly affect the experimental results.

This paper proposed a new deep learning network, called PPM Camouflage Detection Network. The network consists of three main components: Pyramid Pooling Module (PPM), Neighbor Connection Decoder (NCD), and Group Reverse Attention (GRA). Therefore, it is called PNGNet. With respect to dealing with the contradiction between the receptive field of feature extraction and ensuring the resolution of the feature map, the method of fusing the features of different receptive fields and the features of sub-regions was adopted. The feature map was extracted from the camouflage image and divided into two branches; one branch was divided into several sub-regions and the channel size was adjusted by 1*1 convolution. Then, the channel size is restored to the original size by bilinear interpolation (CONV), and finally, the other branch is fused with several sub-regions to obtain the segmentation result of the image.

In short, the paper makes contributions in two aspects.

(1) A new deep learning network for camouflage object detection is raised. By using the PPM (Pyramid Pooling Module), the semantic task of the image segmentation of the camouflage object is more accurate.

(2) After repeated experiments, the proposed PNGNet outperformed most models on three representative camouflage detection datasets and has higher performance and robustness.

2. Related work

According to related research, object detection can be divided into three categories: salient object detection (SOD), general object detection (GOD), and camouflage object detection (COD)[3].

2.1 Salient object detection (SOD)

Salient object detection extracts salient regions in an image by simulating human visual characteristics. The main idea is to detect and encode the features in parallel during visual processing, and then integrate the detected features through centralized attention. The salient object can take the image containing the salient object as the negative sample to help the camouflage object detection[4].

2.2 General object detection (GOD)

GOD has a wider range than SOD (both salient objects and camouflage objects can be detected). It identifies objects in pictures through intelligent algorithms and big data training. The basic idea of GOD lays the foundation for camouflage object detection. However, since camouflage objects can almost be completely fused with the background, the boundary dividing camouflage objects and the environment is difficult to distinguish, and the strong contrast required by segmentation methods is hard to get, it becomes the difficulty of GOD[5].

2.3 Camouflage object detection (COD)

Definition of camouflage object. Input an image to be tested, assign confidence $p_i \in [0,1]$ to each pixel i in the image, and p_i represents the confidence that pixel i belongs to the camouflaged object. The higher the p_i value, the greater the probability that the pixel belongs to the camouflaged object within a certain error range[6].

Application value. Besides academic value, COD is of great significance in searching rare species in the natural field, segmenting pneumonia and polyps in the medical field, monitoring locusts in the agricultural field, and searching for camouflage enemies in the military field.

3. Method

3.1 Overall structure

SINet, a currently widely used model for camouflage detection, simulates the receptive field of the human visual system to develop the feature representations extraction ability of the network. As a rule, the deeper the network, the larger the receptive field. Nevertheless, there is a certain gap between the theoretical receptive field and the actual receptive field in the network. More elaborately, the actual receptive field is smaller than the theoretical receptive field, which makes the network unable to effectively integrate global feature information[7]. Based on this, PNGNet is proposed to solve the above-mentioned problem. The extracted feature values are divided into two branches. One branch is divided into several sub-regions and will be integrated with the other branch so that the features of different receptive fields and the features of sub-regions are fused, and the feature representation ability is enhanced.

Figure. 1 shows the overall framework of the PNG model for camouflage detection. Firstly, the feature map (channel=N) extracted from the input image is pooled to obtain a feature pyramid, which covers from high-resolution low-semantic to low-resolution high-semantic. Specifically, in PNGNet we first input an image of an artifact of size H × W. Features of five levels are extracted from this image, and the resolution of features is

$H/2^k \times W/2^k, k \in \{1,2,3,4,5\}$ respectively. These five levels of features are divided into low-level features { f_k , k = 1, 2} and high-level features { f_k , k = 3, 4, 5}. Previous studies have shown that low-level features cost more to calculate. And it may cause the network to focus more on low-level features and less on high-level features. This means that aggregating low-level features may weaken the experiment effect. Therefore, the PPM module in PNGNet only aggregates high-level features.

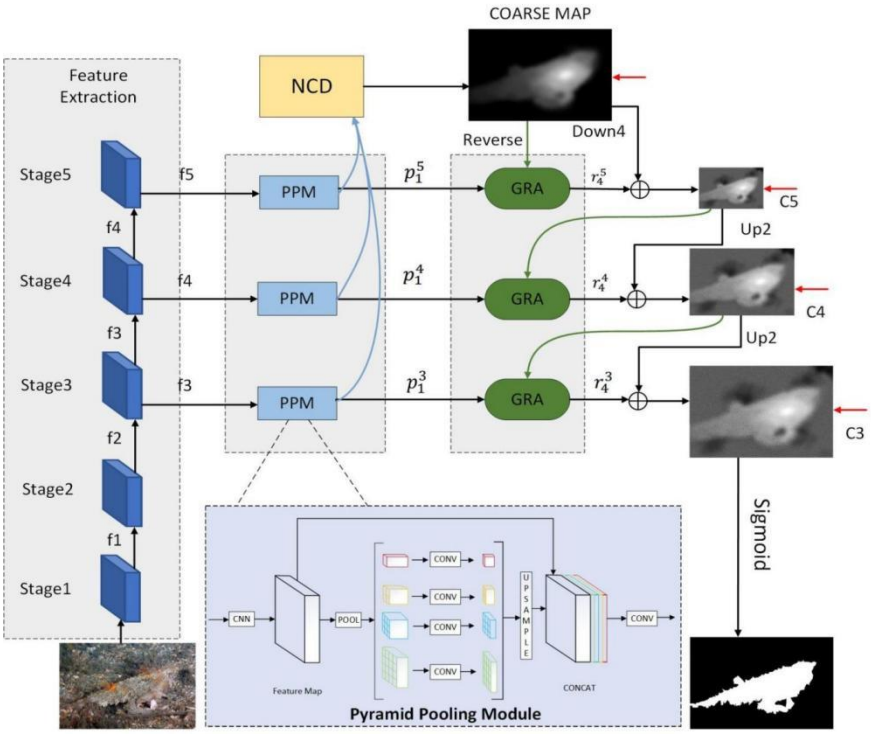

Fig. 1. Flow chart of the PNGNet framework.

Then, the feature map is divided into two branches. One branch is continuously divided into several sub-regions and the channel size is adjusted by 1*1 convolution. Next, the size is restored to the size before being pooled by bilinear interpolation (CONV), and finally, the other branch is fused with several sub-regions to obtain the semantic segmentation prediction result. After the prediction result is obtained, it is passed into a neighbor connection decoder (NCD) to generate a rough positioning map coarse map. In order to refine the structure and texture of the COARSE MAP, the sigmoid function and the reverse operation are used to obtain the reverse guidance of the output, so as to erase the predicted target region in the side output feature and achieve the purpose of extracting complementary regions and details. Finally, in the GRA module, candidate features are fused through group guidance operation, and then the

residual phase is used to generate purified features to obtain a single channel guidance map. Finally, only the optimized guidance map is output as a prediction map for camouflage detection[8]. The key components and loss functions in the PNG model of this article are described in detail below.

PNGNet can be divided into three main parts: pyramid pooling module (PPM), neighbor connection decoder (NCD), and group reverse attention (GRA). PPM is responsible for fusing the features of different receptive fields. NCD is used to locate candidate areas with the assistance of PPM. The GRA module gradually mines more accurate hidden areas by erasing foreground objects.

3.2 PPM vs. NCD

At present, most semantic segmentation frameworks are based on FCN, but FCN is difficult to deal with the relationship and global information between scenes effectively. However, image recognition frameworks mostly rely on pre-trained convolution neural network CNN to aggregate multi-level features. Compared with high-level features, low-level features have less impact on the performance of the deep aggregation method, but due to the high resolution of low-level features, integrating low-level features with high-level features will increase the computational complexity. Inspired by this, in the design of the PNG model, PPM (Pyramid Pooling Module) is used to aggregate the context of different regions to obtain the global context, and a new parallel partial decoder component NCD is proposed to achieve fast and accurate camouflage object detection and obtain more efficient learning ability.

The process is: The feature map (channel =N) is extracted from the input image and pooled to obtain the feature pyramid. 1×1, 2× 2, 4 × 4, 6 ×6 feature maps with channel = 1/N are obtained respectively through 1×1 deep convolution descending channel. Bilinear interpolation was done on the Feature map to fill the unpooled size. To obtain the Feature map with the number of channels doubled, the channel splicing was done. Then, a 1×1 convolution kernel was used for deep convolution of the above Feature map to drop the channel. The prediction result of semantic segmentation is consistent with the number of Feature channels in the input Feature map. After the prediction result is obtained, transmit it to the NCD module of the neighbor connection decoder. Due to the high resolution of low-level features, integrating low-level features with high-level features will increase the computational complexity. Based on the practical application of Fan et al., this paper selectively fuses the features of the highest three layers (i.e.: $f_k \in \mathbf{R}^{W/2^k \times H/2^k \times C}$, $k = 3,4,5$) to achieve higher learning efficiency. In order to keep the semantic consistency within the layer and bridge the context content between, we use the neighbor connection decoder (NCD). The NCD module provides the location information and aggregates the high-level features in a

parallel-joining way[2]. Finally, it produces a rough prediction map, which serves as the global guidance information in the GRA module.

3.3 GRA module

As described in 3.2 of this paper, Coarse Map in this paper comes from the deepest convolution neural network layer and can only predict the relatively rough location of the camouflage without structural details (as shown in Fig.. 1). To solve this problem, we gradually detect the details of hidden areas by erasing the foreground objects in the global prediction map. In particular, we can adaptively learn the reverse attention mechanism in three parallel high-level features, and sequentially explore the complementary regions and details by erasing existing estimated camouflage regions from the high-level side output features[9]. Finally, with the help of this mechanism, several GRA modules are combined to refine the inaccurate and rough prediction regions into accurate and complete prediction maps.

The process is as follows: multiplying the feature { f_k, k=3, 4, 5} output by the high-level side by the reverse attention weight to obtain the output reverse attention feature R_k. The candidate features { $p_i^k \in R^{H/2^k \times W/2^k \times C}$, k=3, 4, 5} are then segmented into m_i=C $/ g_i$ group in the feature dimension, where i=1, 2, 3, g_i represent the group size of the processed features. The attention feature R_k is then periodically inserted into the segmented feature, $p_{i,j}^k \in \mathbf{R}^{H/2^k \times W/2^k \times g_i}$, where i $\in$ {1, 2, 3}, j $\in$ {1, . . . , m_i }, k $\in$ {3, 4, 5}. Finally, the $p_{i,j}^k$ and R_k are fused to generate purified features p_{i+1}^k, and a single-channel prediction map is obtained. Then parametric learning of single-channel graph is carried out by weight W_{GRA}^v. By combining multiple GRA modules, the optimal result is finally obtained.

3.4 Loss function

The loss function is defined as $L = L_{IoU}^W + L_{BCE}^W$, where L_{IoU}^W and L_{BCE}^W represents weighted IoU loss and binary cross entropy (BCE) loss based on global and local pixel level constraints, respectively[10].

Standard IOU losses are commonly used in image segmentation tasks, but in PNGNet we use weighted IOU losses instead of standard IOU losses, which can refine the camouflage object region more accurately by increasing the weight of difficult pixels. In the same way, the standard binary cross-entropy loss is not used. Instead, the more difficult pixels are emphasized. The validity of these loss function definitions has been demonstrated in the field of significant target detection. Here, the three outputs and the Coarse Map are under deep supervision,

and each prediction map is up sampled to the same size as the truth value figure G_s.

4. Details of experiment

4.1 Learning & training strategy

The loss function $L = L_{IoU}^{W} + L_{BCE}^{W}$ mentioned in Section 3.4 is used to calculate global and local (pixel-level) constraints. In the field of image segmentation, the standard IOU loss has been widely used. In order to highlight the importance of difficult pixels, the weighted cross-Union ratio loss is introduced. Similarly, in this paper, more attention is paid to the more difficult pixels than to the standard binary cross-entropy loss, rather than giving each pixel the same weight. In addition, depth supervision (i.e., C3, C4, and C5) is used for the three by-pass outputs and the global feature COARSE MAP. Each prediction map is upsampled to the same size as the truth value map G. Therefore, the complete loss function of the PNGNet model is defined as:

$$\mathrm{L}_{\text{total}} = \mathrm{L}_{seg}(G_s, S_g^{up}) + \mathrm{L}_{\text{edge}} + \sum_{i=3}^{i=5} \mathrm{L}_{seg}\left(G_s, S_g^{up}\right) \tag{1}$$

4.2 Dataset

This paper uses the largest camouflage object detection dataset COD10K (a challenging, high-quality, densely labeled dataset) to train and test data. The CAMO and COD10K datasets, which are also widely used in the field of counterfeit detection, are also adopted for testing. The CAMO dataset includes 250 pictures covering 8 categories. The CHAMELEON dataset has 76 Truth Value Maps (GTS) with manually labeled target levels, collected by Google's search engine using "hidden animals" as a search keyword. The COD10K dataset consists of 6066 pictures divided into 10 parent groups: flying animals, aquatic animals, terrestrial animals, amphibians, other groups, skies, vegetation, indoor, marine, and sand. The 10 groups can be further divided into 78 subgroups: 69 hidden groups and 9 non-hidden groups.

Most of the hidden images are from Flickr and are limited to academic research purposes, while the rest are from other websites, namely: Visual Hunt, Pixabay, Unsplash, Free Images, and other sites for publishing archived images in the public domain without copyright or ownership issues. With COD10K labeled by Fan et al., the largest COD dataset with the richest tags was constructed so far. This article draws on the work of Fan et al. in the selection of COD datasets. In order to provide a large number of training data for the deep learning model, the COD10K dataset is randomly divided into a training dataset of 4040 images and a test dataset of 2026 images. The CAMO and CHAMELEON datasets were used as test datasets to verify the generalization ability of PNGNet.

4.3 Implementation Details

Our parameter setting followed the work of Fan et.al. During the training stage, the batch size is set to 36, and the learning rate starts at 1e-4. Our PNGNet is implemented in PyTorch and trained with the Adam optimizer. The running time is measured on two GeForce RTX 2070. The input images are all adjusted to the size of 352×352, which is also consistent with the setting of Fan et al.

4.4 Evaluation index

With regard to the mean absolute error (MAE) proposed by Perazzi et al. in the field of SOD, this paper uses the mean absolute error M index (MAE(M)) to evaluate the pixel-level accuracy between the predicted image C and the true image G[11]. MAE is widely used for COD tasks, but it can not determine where errors occur. This paper refers to an enhancement-matching evaluation index (E-measure, E_ϕ) based on the mechanism of human visual perception proposed by Fan et al. as the evaluation standard, which takes into account both the matching of pixel-level information and the statistics of image-level information. Since the shape, color, and size of all kinds of camouflage are different, it is difficult to test. This paper also adopts the S evaluation measure (S-measure, S_α) and weighted F evaluation measure (F-measure, F_β^w) as evaluation indexes, which are specifically expressed as follows.

Structural index (S_α): This index measures the results from the perspective of the human visual system. It is used to measure the structural similarity between the prediction map and the truth map:

$$S_\alpha = (1-\alpha) * S_o(S_p, G) + \alpha * S_r(S_p, G) \tag{2}$$

In this formula, α is a balance coefficient for controlling the object-level similarity S_o and the region-level similarity S_r [12].

E_ϕ Mean: This is a recently proposed indicator that measures both the local and global similarity between two binary maps simultaneously. The formula is listed below:

$$E_\phi \text{ mean} = \frac{1}{w \times h} \sum_{x}^{w} \sum_{y}^{h} \varnothing(S_p(x, y), G(x, y)) \tag{3}$$

In this formula, w and h represent the width and height of the truth-value map G, and (x, y) represents the coordinates of each pixel in G. The symbol ϕ is an enhanced alignment matrix. The prediction graph S_p is threshed with a threshold of 0 to 255 to obtain a set of binary graphs so that a set of E_ϕ scores can be obtained. In this experiment, the mean of E_ϕ at all thresholds is reported.

Mean absolute error (MAE): This index measures the pixel-level error between S_p and G, and is defined as the following formula:

$$\mathrm{MAE} = \frac{1}{w \times h} \sum_{x}^{w} \sum_{y}^{h} \left| S_p(\mathrm{x}, \mathrm{y}) - \mathrm{G}(\mathrm{x}, \mathrm{y}) \right| \tag{4}$$

F-measure(F_β^w): F-measure indicates the effectiveness of a test method. The higher the F-measure, the more reliable the test method. The formula is as follows:

$$F_\beta^w = (1+\beta^2) \frac{\mathrm{Pr}ecision^w \cdot \mathrm{Re}call^w}{\beta^2 \cdot \mathrm{Pr}ecision^w + \mathrm{Re}call^w} \tag{5}$$

As stated by R. Margolin et al., this formula can better measure the overall results[13]. F_β^w adds weighted precision, which can deal with the problem of equally important defects. This weighted F-measure is more effective than the traditional F_β.

5. Results of the experiment

5.1 Qualitative results

Compared with SINet, PNGNet has a further improvement in visual performance under different lighting conditions, appearance changes, and blurred boundaries.

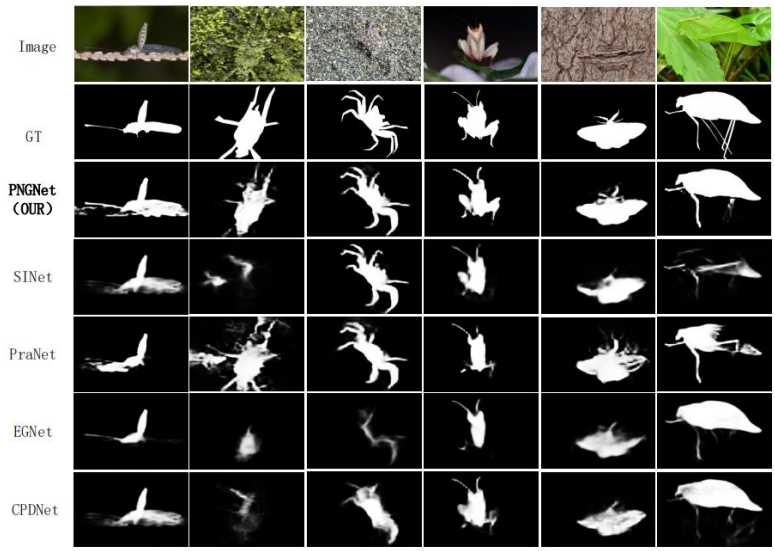

Fig. 2 Performance comparison between PNGNet and other models.

Although SINet can position hidden targets, the prediction results are still not accurate enough. In this paper, the PNGNet structure enhances the ability of feature extraction by fusing the features of different receptive fields and sub-

regions in the image segmentation stage. For challenging cases (as shown in Fig. 2), PNGNet can predict the correct hidden target and its details more accurately than SINet, showing the robustness of the network structure.

5.2 Quantitative results

To quantitatively compare the performance of the two models for camouflage detection, we present the quantitative results in Table 1. As can be seen from Table 1,

Table 1

Quantitative results on three datasets

Models	CAMO_TestingDataset				CHAMELEON_TestingDataset				COD10K_TestingDataset			
	Sα↑	$E_\phi \uparrow$	$F_w^\beta \uparrow$	M↓	Sα↑	$E_\phi \uparrow$	$F_w^\beta \uparrow$	M↓	Sα↑	$E_\phi \uparrow$	$F_w^\beta \uparrow$	M↓
PNGNet (OUR)	0.807	0.864	0.725	0.077	0.825	0.882	0.745	0.066	0.813	0.884	0.685	0.041
SINet	0.790	0.847	0.708	0.080	0.814	0.871	0.735	0.068	0.812	0.883	0.684	0.041
SINet (FAN)	0.869	0.891	0.740	0.044	0.751	0.771	0.606	0.100	0.771	0.806	0.551	0.051
FPN	0.794	0.783	0.590	0.075	0.684	0.677	0.483	0.131	0.697	0.691	0.411	0.075
PFANet	0.679	0.648	0.378	0.144	0.659	0.622	0.391	0.172	0.636	0.618	0.286	0.128
HTC	0.517	0.489	0.204	0.129	0.476	0.442	0.174	0.172	0.548	0.520	0.221	0.088
PraNet	0.860	0.907	0.763	0.044	0.769	0.824	0.663	0.094	0.789	0.861	0.629	0.045
PSPNet	0.773	0.758	0.555	0.085	0.663	0.659	0.455	0.139	0.678	0.680	0.377	0.080
UNet++	0.695	0.762	0.501	0.094	0.599	0.653	0.392	0.149	0.623	0.672	0.350	0.086
PiCANet	0.769	0.749	0.536	0.085	0.609	0.584	0.356	0.156	0.649	0.643	0.322	0.090

PNGNet in this paper is better than SINet in terms of S_α, E_ϕ mean, F-measure, and MAE indicators. The main reason for the significant performance improvement of segmentation results is that in PNGNet, we aggregate the contexts of different regions through PPM to obtain the global context, which provides the expression of robustness. We also introduce a semi-supervised learning strategy into PNGNet to further improve the performance of the Dice index. As a camouflage detection tool, the model shows good quantitative evaluation results on each data set. On the CHAMELEON dataset, as can be seen from Table 1, PNGNet exceeds the performance of the SINet model on all indicators. We also tested performance on the CAMO dataset, which contains a variety of hidden targets and is more challenging than the CHAMELEON dataset. PNGNet still outperforms the SINet model across the board, showing the robustness of the model. On COD10K, the largest camouflage object detection data set to date, we observe that PNGNet is still superior to SINet, which shows that PNGNet has a stronger ability to deal with feature aggregation than the SINet model and can learn richer and more diverse features from rough to fine, which makes an important contribution to accurately identifying the boundary between camouflage objects and environment. Overall, based on the various metrics

mentioned in Section 4.3 of this article, PNGNet outperforms SINet in camouflage recognition across the board.

The best score is identified in bold. ↑ represents the higher the score (the better). E_ϕ represents the average of the E index. Note: The data in the third to last lines refer to Fan's test results [2], and the SINet test results are different from the results we actually ran.

5.3 Generalization

This model uses part of COD10K images as the training set. The CAMO, CHAMELEON datasets, and the rest of COD10K images were used as test datasets. As shown in Table 1, our model still performs better than SINet in these two datasets. Additional datasets were introduced to demonstrate the strong generalization ability of PNGNet. Although our model performed well on other datasets CAMO and CHAMELEON, we did not re-train.

5.4 Ablation Studies

PNGNet aggregates three high-level feature maps distributed at Stage3, Stage4, and Stage5. Low-level feature maps contribute little to the performance of the deep integration model and have a relatively large spatial resolution. If low-level feature maps are aggregated together, the calculation consumption will increase. Therefore, we conducted experiments to explore the effectiveness of aggregation of three high-level feature maps.

Table 2

Comparison of feature aggregation strategies

Models	CAMO _Testing Dataset				CHAMELEON _Testing Dataset				COD10K_Testing Dataset			
	Sα↑	E_ϕ ↑	F_w^β ↑	M↓	Sα↑	E_ϕ ↑	F_w^β ↑	M↓	Sα↑	E_ϕ ↑	F_w^β ↑	M↓
Stage3	0.742	0.833	0.707	0.092	0.781	0.823	0.715	0.078	0.779	0.849	0.643	0.060
Stage3+Stage4	0.788	0.848	0.720	0.081	0.817	0.877	0.738	0.069	0.808	0.877	0.667	0.044
Stage3+Stage4+Stage5	**0.807**	**0.864**	**0.725**	**0.077**	**0.825**	**0.882**	**0.745**	**0.066**	**0.813**	**0.884**	**0.685**	**0.041**

As shown in Table 2, 'Stage3' means that only f3 is associated with the PPM module. "Stage3+Stage4" means that only f3 and f4 feature maps are aggregated. 'Stage3+Stage4+Stage5' represents the aggregation of all high-level feature maps, which is the method adopted by PNGNet. Stage 3 and Stage3+Stage4 have significant differences in these four indicators (Sα , Eφ , M, F_β^w) compared with Stage3+Stage4 and Stage3+Stage4+Stage5. This suggests that higher-level feature maps truly contribute to better results, while little improvement of Stage3+Stage4+Stage5 illustrates that the current higher-level feature maps have gone up to the limit and is the best result for the experiment.

6. Conclusion

This paper proposes a new deep neural network PNGNet for concealed object detection tasks, which surpasses SINet proposed by Fan et al. on

CHAMELEON, CAMO, and COD10K datasets. In the stage of feature extraction, PNGNet deals with the contradiction between the receptive field of feature extraction and the resolution of the feature map. Specifically, the feature image was extracted from the camouflage image and divided into two branches. One branch was divided into several sub-regions and the channel size of the sub-region was adjusted. Then, the channel size was restored to the original size and the other branch was fused with it. Therefore, the features of different receptive fields and sub-regions are fused, and the feature representation ability is enhanced. This provides a new idea for concealed object detection. In the stage of feature extraction, the pyramid pooling module can effectively integrate global feature information, which makes the semantic task of image segmentation of camouflaging objects more accurate. In the stage of identification, PNGNet adopts the progressive strategy on the basis of SINet to effectively maintain the generalizability of the model. In general, PNGNet, with competitiveness, can achieve ideal results visually and can predict the correct hidden target and its details more accurately than SINet, showing the robustness of the PNGNet structure. In addition to the value in the field of camouflage detection, the results are of great significance in searching rare species in the natural field, segmenting pneumonia and polyps in the medical field, monitoring locusts in the agricultural field, and searching camouflage enemies in the military field.

REFERENCES

[1] *D. P. Fan, G. P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao,* "Pranet: Parallel reverse attention network for polyp segmentation", in Med. Image. Comput. Comput. Assist. Interv., 2020.

[2] *D. P. Fan, G. P. Ji, G. L. Sun, M. M. Cheng, J. B. Shen, L. Shao,* "Camouflaged Object Detection", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2777-2787.

[3] *Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu,* "Object detection with deep learning: A review", IEEE T. Neural Netw. Learn. Syst., **vol.** 30, no. 11, 2019, pp. 3212–3232.

[4] *F. Perazzi, P. Kr¨ahenb¨uhl, Y. Pritch, and A. Hornung,* "Saliency filters: Contrast based filtering for salient region detection", in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. IEEE, 2012, pp. 733–740.

[5] *M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman,* "The PASCAL visual object classes challenge: A retrospective", Int. J. Comput. Vis., **vol.** 111, no. 1, 2015, pp. 98–136.

[6] *Q. Zhai, X. Li, F. Yang, C. Chen, H. Cheng, and D. P. Fan,* "Mutual graph learning for camouflaged object detection", in IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[7] *H. S. Zhao, J. P. Shi, X. J. Qi, X. G. Wang, and J. Y. Jia,* Pyramid scene parsing network", In CVPR, 2017, pp. 2881–2890.

[8] *X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand,* "Basnet: Boundary-aware salient object detection", in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 7479–7489.

[9] *N. Xu, B. Price, S. Cohen, and T. Huang,* "Deep image matting", in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 2970–2979.

[10] *J. Wei, S. Wang, and Q. Huang,* "F3Net: Fusion, Feedback and Focus for Salient Object Detection", in AAAI Conf. Art. Intell., 2020.

[11] *F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung,* "Saliency filters: Contrast based filtering for salient region detection", in IEEE Conf. Comput. Vis. Pattern Recog., 2012, pp. 733–740.

[12] *D. P. Fan, M. M. Cheng, Y. Liu, T. Li, and A. Borji,* Structure-measure: A New Way to Evaluate Foreground Maps", in Int. Conf. Comput. Vis., 2017, pp. 4548–4557.

[13] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?", in IEEE Conf. Comput. Vis. Pattern Recog., 2014, pp. 248–255.