FocusNet: Cascaded Lightweight Networks and Ascending Feature Enhancement for Efficient Salient Object Detection

Chiheng Zhou[1], Yongxia Zhou[1]✉, and Chen Pan[1]

China Jiliang University, Xueyuan Street, 310000, Zhejiang Hangzhou, China
{s22030812011,zhouyongxia,pc916}@cjlu.edu.cn

Abstract. Existing salient object detection methods typically depend on large, pretrained backbone networks for feature extraction. While this enhances performance, their large size and high number of parameters make them less practical for widespread use in real-world applications. In contrast, lightweight backbone networks are smaller and have fewer parameters, yet they typically fall short in delivering strong feature extraction and precise localization. To address this issue, this paper introduces FocusNet, an innovative lightweight solution. By cascading lightweight networks through the FOCUS module, we simulate the human eye's focusing process, dividing localization and feature extraction into two independent and sequential steps, significantly enhancing the feature extraction capabilities. Additionally, utilizing the Ascending Feature Enhancement (AFE) strategy, we progressively saturate the localization of salient objects from shallow to deep layers, significantly improving localization accuracy. Extensive experiments on five public datasets demonstrate that our method, while maintaining a low parameter(3.15M), performs comparably to methods based on large backbone networks. . The source code will be released at https://github.com/YinYinOvO/FocusNet.

Keywords: Salient Object Detection · Efficient · Feature Enhancement

1 Introduction

Salient object detection[1] is a fundamental task in computer vision, aimed at identifying and segmenting the most attention-grabbing areas in images. This technology plays a crucial role in fields such as visual tracking[2], semantic segmentation[3, 4], video processing[5], and image retrieval[6]. With the rapid advancement of deep learning, neural network-based methods for salient object detection have become mainstream. These methods typically use deep semantic information to locate salient objects and refine their boundaries with shallow detail features, significantly improving detection performance.

Despite the significant progress made by deep learning approaches in salient object detection, the high precision often comes at the cost of substantial computational expenses and large model sizes, limiting their application in resource-constrained environments, especially on mobile devices. To address the lightweight

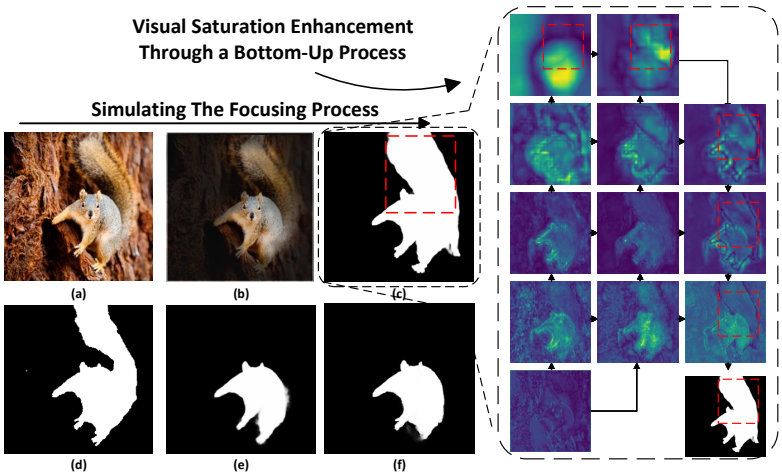

Fig. 1. Overview of the method we proposed. (a) shows the original image; (b) simulates the human eye's focusing process; (c) presents our prediction results; (d) displays the ground truth; (e) and (f) compare predictions from EDN[7] and MENet[8]. The visualization of our feature enhancement method is featured on the right.

challenge in salient object detection, many studies have attempted to design independent feature extractors tailored for this task, such as CSNet[9] and U^{2}Net[10], or have adopted lightweight backbone networks like MobileNets[11], ShuffleNets[12], and MobileVits[13] in place of traditional heavy backbones like ResNet[14] and Swin-Transformer[15].

While these studies have achieved some success in lightweight design, they often compromise performance: fully independent feature extractors, though flexible in design, lack the semantic priors from backbone networks, resulting in poorer performance and slower convergence; lightweight backbone networks reduce computational load but typically fall short in feature extraction and target localization capabilities.

Inspired by the human visual system, this paper presents the FOCUS module, which emulates the process of human eye focus: first rapidly scanning the scene to locate an object of interest, and then precisely focusing on it to enhance feature extraction. This sequential approach leverages a cascade of two lightweight backbone networks, where the first network handles the localization of salient objects, and the second network refines feature extraction based on the localized information. This design effectively reduces feature redundancy and significantly improves the precision of detailed segmentation.

Additionally, we have developed a ascending feature enhancement method(AFE) that is different from the top-down process in the decoding module of traditional methods. Instead, it amplifies the commonalities between adjacent stage features

through a bottom-up process, enhances the localization of foreground targets, and uses a gradually saturated form to enhance the localization of salient targets, effectively enhancing the accuracy of localization. As shown in 1, the AFE method enables accurate segmentation of areas that were previously prone to under-segmentation.

Extensive experiments demonstrate that the FOCUS module significantly enhances the feature extraction capabilities of lightweight backbone networks, allowing methods based on lightweight networks to perform comparably to those based on ResNet50. Our ascending feature enhancement method leverages the class-agnostic properties of salient object detection to significantly improve performance and localization abilities with almost no increase in the number of parameters.

The main contributions of this paper include:

- We proposed the FOCUS module, which optimizes the localization and feature extraction processes for salient object detection by simulating the human eye's focusing mechanism.
- We developed a bottom-up feature enhancement method (AFE) that effectively improves the performance of lightweight networks in salient object detection.
- Extensive experiments on five public datasets have demonstrated the effectiveness of our method, enabling methods based on lightweight backbone networks to reach a level comparable to those of large-scale networks.

2 Related Work

2.1 Salient Object Detection

With the rapid advancement of deep learning, neural network-based methods for salient object detection have significantly progressed beyond traditional techniques. Modern approaches largely focus on enhancing performance through the design of efficient feature fusion modules, utilizing edge information, and developing effective loss functions. For instance, methods like PoolNet[16], MINet[17], and F3Net[18] have optimized the use of features across different levels by implementing efficient multi-scale feature fusion modules. Techniques such as EGNet[19] and RCSB[20] concentrate on using edge information to guide the precise segmentation of salient objects. Furthermore, the introduction of effective loss functions like SSIM[21] Loss and edge-weighted loss in methodologies such as Contour-aware Loss[22], TRACER[23], and BASNet[24] has significantly improved model performance.

2.2 Efficient and Lightweight Salient Object Detection

Lightweight salient object detection has consistently been a topic of interest. Methods like CSNet[9] and U^{2}Net[10] have opted to bypass pretrained backbone networks, instead developing efficient feature extraction modules specifically optimized for salient object detection, which substantially reduce the overall model size. However, the absence of semantic priors typically provided by

pretrained backbones often results in these methods facing challenges with poor performance and difficult training. On the other hand, techniques such as MobileSal[25] incorporate lightweight backbone networks like MobileNet[11] and ShuffleNet[12], effectively reducing model size. Yet, these lightweight networks generally exhibit weaker performance compared to their heavier counterparts. Consequently, our focus has been on how to elevate the feature extraction capabilities of lightweight networks to match those of heavier backbone networks.

2.3 FPN and Its Variants

Feature Pyramid Networks (FPN)[26] and their variants highlight the importance of feature fusion across different layers. Subsequent studies, such as PANet[27], have refined this architecture using a bottom-up and then top-down connection strategy. NAS-FPN[28] discovered an irregular connection structure through architectural search, while EfficientDet[29] built on this foundation to further optimize the feature fusion process by repeatedly using the decoding module to enhance result accuracy. Our approach, although structurally similar to BiFPNs, diverges completely in purpose and functionality. BiFPNs primarily focus on repeatedly fusing extracted features, whereas our FOCUS module selectively extracts features guided by location information, preventing the dilution of deep-layer information by shallow layers. Moreover, our AFE method performs saturated feature enhancement from bottom to top, avoiding the problem of significant differences in response between different salient regions in deep features.

3 Proposed Method

In this section, we will provide a detailed description of the methodology we have proposed to effectively address the limitations in feature extraction and target localization capabilities of lightweight backbone networks.

3.1 FOCUS Module

Inspired by the human visual mechanism for capturing objects of interest, we designed the FOCUS module to compensate for the deficiencies in feature extraction capabilities of lightweight backbone networks. In the FOCUS module, we employ two identical lightweight backbone networks, transforming the originally integrated processes of feature extraction and localization into two independent and sequential processes. We refer to the first network as the localization network, which coarsely locates salient targets, allowing the second feature extraction network to selectively extract features guided by this localization information. To ensure accurate priors from the localization network, we added a auxiliary localization loss at its end and used the Feature Refine Module (FRM) to transmit localization information to the superficial layers, guiding the feature extraction network to effectively extract features in the early stages. The

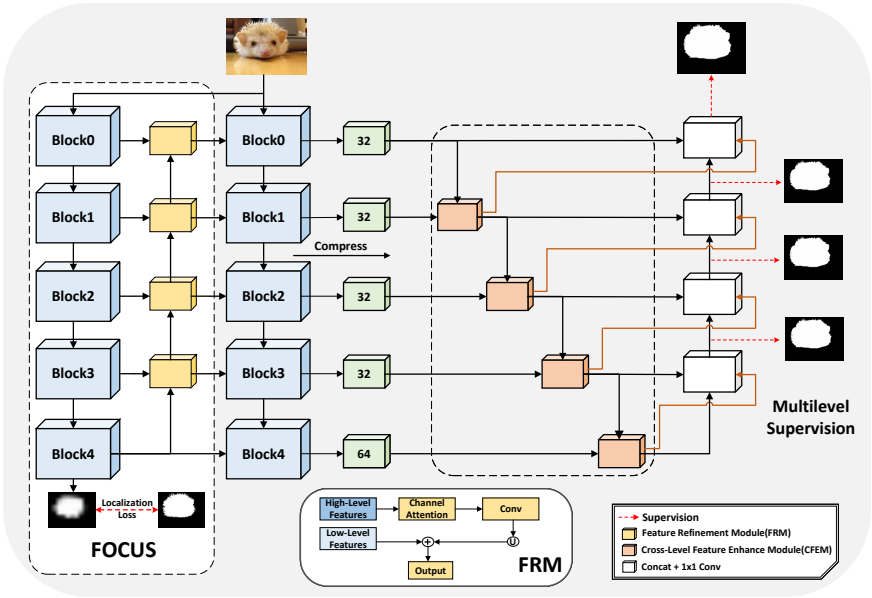

Fig. 2. This diagram illustrates the overall architecture of our model, with a detailed implementation of the Feature Refine Module(FRM) shown below.

structure of the FRM, shown in 2, involves up-sampling the deep features after channel weighting and adjusting their channel numbers through a 1x1 convolution before combining them with the shallow features.

To demonstrate that the performance improvements brought by the FOCUS module are not merely due to increased model complexity, we presented three different network combinations in 3. The results on the left side of 3 clearly indicate that simple cascading of backbone networks increased model complexity and computational costs but did not significantly enhance performance. With the introduction of auxiliary localization loss, the responsibilities of the two networks were clarified, significantly boosting performance; further transmitting localization information to the superficial layers led to additional performance gains. This confirms that the FOCUS module improves performance by altering the traditional paradigm of feature extraction, not just by increasing complexity.

To further prove its effectiveness and generalizability, we conducted experiments on various backbone networks of different structures and sizes. Significant performance improvements were observed on lightweight backbone networks such as MobileNetv2[11], EfficientNet-b0[30], and MobileVit[13], while not very significant enhancements were seen on larger backbone networks like ResNet50[14] and Swin-Base[15], which also added a substantial amount of extra parameters and computational costs. We speculate that this is because the feature extraction capabilities of large backbone networks are already sufficient to handle localization

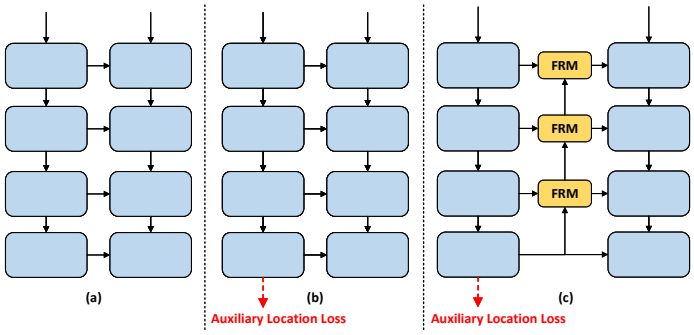

Fig. 3. This figure displays different methods of combining backbones: (a) depicts a simple horizontal concatenation, (b) introduces auxiliary localization loss, and (c) builds on the previous by incorporating the FRM to transfer deep-layer information to shallow layers.

and feature extraction within a single network. To validate this hypothesis, we independently fed features extracted by the two networks into the decoding module. Results indicated that for heavy backbone networks, the localization network alone was accurate enough for salient object segmentation, whereas the feature extraction network did not perform as expected. In contrast, for lightweight backbone networks, neither the localization nor the feature extraction network alone was sufficient for effective segmentation; only their combination achieved optimal segmentation results.

3.2 Ascending Feature Enhancement(AFE)

Incomplete segmentation of salient objects has always been a focal issue in this field. Previous methods, such as PoolNet[16], have attempted to address this by designing a Global Guidance Module that feeds deep-layer information back to shallow layers, while PA-KRN[31] employs a knowledge recall approach to combine the final prediction map with features from different stages. Through visualizing features at various stages, we have founded the significant cause of this issue: deep features exhibit uneven response intensities to multiple salient objects or different regions of a single object, leading to disturbances in weaker response areas when integrating deep features with shallow features, which can result in incomplete final segmentation. The visualization further reveals that the feature representation becomes progressively abstract from shallow to deep layers, with features belonging to the foreground being enhanced and those of the background being suppressed.

Leveraging the class-agnostic nature unique to salient object detection tasks, we have developed an Ascending Feature Enhancement method that saturatively enhances the feature response of various foreground areas. The differences be-

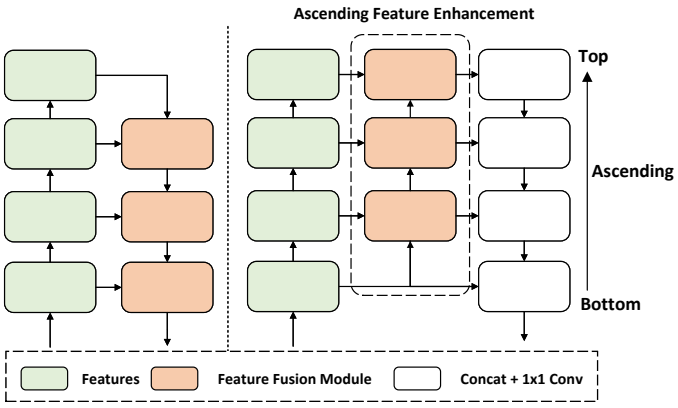

Fig. 4. This diagram highlights the differences in the decoding process between our Ascending Feature Enhancement method and traditional approaches.

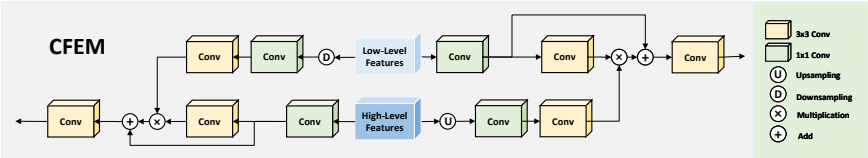

Fig. 5. This diagram illustrates the detailed implementation of the Cross-layer Feature Enhancement Module (CFEM).

tween it and other methods in the decoding part can be clearly observed in 4.

The Ascending Feature Enhancement (AFE) consists of multiple Cross-layer Feature Enhancement Modules (CFEM), designed to amplify the commonalities between features across adjacent stages, thereby strengthening the feature response in foreground areas. The specific implementation of CFEM is shown in 5. Assuming that s1, s2, s3, s4, and s5 are the features extracted at various stages by the feature extraction network, we proceed with further processing and enhancement based on these features.

$$s_1^{deep}, s_2^{shallow} = CFEM_1(s_1, s_2) \tag{1}$$

$$s_i^{deep}, s_{i+1}^{shallow} = CFEM_i(s_i^{shallow}, s_{i+1}), i = \{2, 3, 4\} \tag{2}$$

Aside from s_1 and s_5, each stage features two enhanced features of the same scale(s_i^{deep},$s_i^{shallow}$,i=$\{2,3,4\}$). These enhanced features are concatenated with up-sampled deep features, followed by simple fusion through a 1x1 convolution.

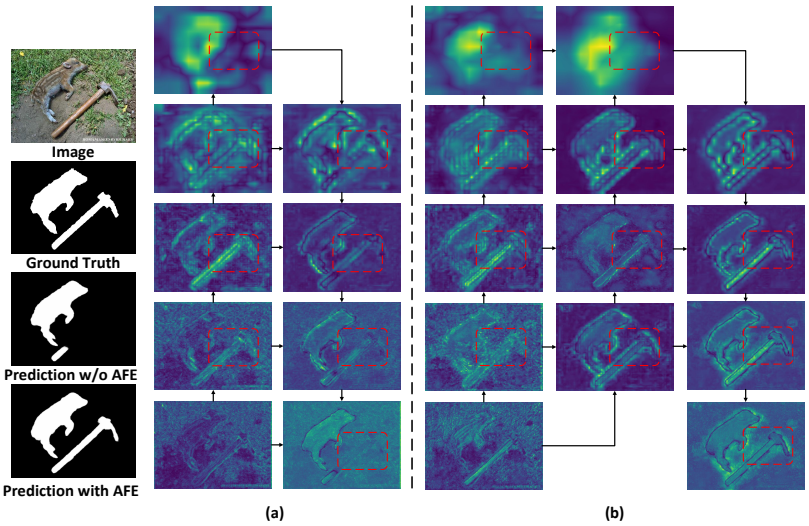

Fig. 6. This figure demonstrates how the Ascending Feature Enhancement (AFE) method enhances localization information within deep layers by altering feature flow in the decoding module, effectively addressing issues with incomplete segmentation.

Subsequently, the fused features are combined with s_1^{deep} to produce the final prediction map, which is then up-sampled to the size of the input.

With the aid of the CFEM, the AFE modifies the traditional feature flow during the decoding process, shifting from a top-down to a bottom-up approach. This change amplifies the commonalities between adjacent stages, effectively achieving precise relocalization of salient targets. 6 provides a visualization of this process, clearly demonstrating how AFE significantly enhances the integrity of salient objects.

3.3 Loss Function

Our loss function is divided into two parts: the auxiliary localization loss at the end of the localization network and the precise segmentation loss during the decoding process.

$$L_{Total} = L_{AuxiliaryLocalizationLoss} + L_{PreciseSegmentationLoss} \tag{3}$$

For the auxiliary localization loss, we downscale the mask to match the size of the deep features and perform a coarse localization of potential salient regions using a straightforward BCE loss. we employ a composite loss function comprising a weighted BCE, weighted IoU, and a weighted L1 loss, details of which are elaborated in TRACER[23].

$$L_{PreciseSegmentationLoss} = w1 \times L_{BCE} + w2 \times L_{IoU} + w3 \times L_{L1} \tag{4}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluated our model on five popular benchmark datasets: HKU-IS[32] with 4447 images, DUTS[33] with 10,553 images for training and 5,019 images for testing, ECSSD[34] with 1,000 images, PASCAL-S[35] with 850 images and DUT-OMRON[36] with 5,168 images. For the evaluation metrics, we used mean absolute error (MAE), weighted F-measure and S-measure[37]. We trained FocusNet on DUTS-TR[33] dataset and tested on the other datasets.

4.2 Implementation Details

In the training phase, we enhance the generalization ability of the model through data enhancement methods such as horizontal flipping, rotation and blurring. Initialize the backbone network using weights pre-trained on the ImageNet dataset. We set the batchsize to 16 and the epoch to 100. We use the Adam optimizer and set the learning rate to 1×10^{-4} and the weight decay to 10^{-4}. For fair comparison, we fixed all random seeds at 42. Our model is built on Pytorch 1.8.1 and a single 3090 GPU (with 24GB of memory).

4.3 Comparison with State-of-the-arts

We compared our model with 12 state-of-the-art methods. To be fair, the results of all 12 methods were tested under the same evaluation code directly using the salient prediction maps provided by the authors and without any post-processing.

Quantitative Comparison As presented in 1, we conducted a comparative analysis of our proposed models, FocusNet and FocusNet+, against other state-of-the-art approaches utilizing ResNet50[14] and MobileNetv2[11] as backbone architectures. FocusNet+ denotes a model that has been equipped with the FOCUS module. The quantitative results reveal that even in the absence of the FOCUS module, our approach, augmented by the other modules, already exhibits robust competitiveness. With the incorporation of the FOCUS module, there is a further enhancement in performance, achieving optimal or near-optimal results across various metrics.

While a performance surge is evident, the increment in parameter count for the ResNet50[14] backbone is substantial. Consequently, in terms of performance gain relative to increased complexity, it is not the most efficient trade-off. Our method based on MobileNetv2[11], however, demonstrates performance comparable to many ResNet50-based methods with only a marginal increase in parameters, providing a novel perspective for enhancing the performance of lightweight models.

Table 1. Quantitative comparisons with 12 state-of-the-art methods on five datasets with MAE,weighted F-measure and S-measure. FocusNet+: Our method with FCOUS module.(The best results are highlighted in red and the second-best in blue.)

				ECSSD			HKU-IS		
Method	Year	Backbone	Params	MAE↓	wF↑	Sm↑	MAE↓	wF↑	Sm↑
PoolNet[16]	2019	ResNet50	68.3	.0388	.8962	.9208	.0321	.8830	.9167
CPD[38]	2019	ResNet50	47.9	.0371	.8980	.9181	.0342	.8747	.9055
BASNet[24]	2019	ResNet50	87.1	.0370	.9038	.9162	.0322	.8891	.9089
F3Net[18]	2020	ResNet50	25.5	.0333	.9122	.9242	.0280	.9004	.9172
MINet[17]	2020	ResNet50	162.4	.0335	.9107	.9249	.0285	.8973	.9188
U^2Net[10]	2020	RSU	44.0	.0330	.9102	.9276	.0312	.8891	.9158
SGL-KRN[31]	2021	ResNet50	68.9	.0360	.9101	.9231	.0280	.9042	.9206
PA-KRN[31]	2021	ResNet50	102.2	.0323	**.9184**	.9278	.0271	.9093	**.9235**
ICON[39]	2022	ResNet50	33.1	**.0318**	.9182	**.9290**	.0289	.9022	.9200
RCSB[20]	2022	ResNet50	26.9	.0335	.9158	.9217	.0268	**.9086**	.9187
EDN[7]	2022	ResNet50	42.9	.0320	.9177	.9267	**.0264**	.9077	**.9240**
FocusNet	2024	ResNet50	28.1	.0330	.9160	.9232	.0273	.9062	.9188
FocusNet+	2024	ResNet50	57.2	**.0291**	**.9247**	**.9289**	**.0252**	**.9133**	.9232
CSNet[9]	2020	gOctConv	0.78	.0600	.7770	.8810	.0660	.8060	.8930
U^2Netp[10]	2020	RSU	1.1	.0409	.8854	.9179 .	.0370	.8668	.9079
EDN[7]	2022	MobileNetv2	1.80	.0420	.8898	**.9106**	.0338	.8787	**.9072**
FocusNet	2024	MobileNetv2	1.76	**.0384**	**.8996**	.9093	**.0298**	**.8907**	.9063
FocusNet+	2024	MobileNetv2	3.15	**.0368**	**.9074**	**.9141**	**.0282**	**.8969**	**.9106**

		DUTS-TE			PASCAL-S			DUT-OMRON		
Method	Year	MAE↓	wF↑	Sm↑	MAE↓	wF↑	Sm↑	MAE↓	wF↑	Sm↑
PoolNet[16]	2019	.0403	.8070	.8833	.0745	.7983	.8491	.0555	.7291	.8357
CPD[38]	2019	.0434	.7951	.8690	.0705	.7999	.8481	.0560	.7190	.8248
BASNet[24]	2019	.0476	.8030	.8657	.0750	.7983	.8382	.0565	.7509	.8362
F3Net[18]	2020	.0354	.8349	.8884	.0615	.8218	.8604	.0526	.7473	.8385
MINet[17]	2020	.0372	.8254	.8840	.0635	.8148	.8559	.0555	.7377	.8329
U^2Net[10]	2020	.0448	.8036	.8736	.0736	.7973	.8444	.0545	.7567	.8466
SGL-KRN[31]	2021	.0337	.8470	.8929	.0674	.8183	.8555	**.0492**	.7646	.8464
PA-KRN[31]	2021	.0328	**.8609**	**.9005**	.0660	.8227	.8578	.0496	**.7792**	**.8533**
ICON[39]	2022	.0373	.8366	.8886	.0637	.8237	.8613	.0569	.7606	.8444
RCSB[20]	2022	.0350	.8396	8808	.0592	.8322	.8597	**.0492**	.7524	.8351
EDN[7]	2022	.0351	.8451	.8924	.0614	.8326	**.8644**	.0494	**.7696**	**.8495**
FocusNet	2024	**.0308**	.8520	.8929	**.0581**	**.8338**	.8635	.0512	.7466	.8330
FocusNet+	2024	**.0296**	**.8614**	**.8974**	**.0561**	**.8379**	**.8642**	**.0487**	.7548	.8366
CSNet[9]	2020	.0750	.6440	.8220	.1040	.6910	.8140	.0800	.6200	.8050
U^2Netp[9]	2020	.0545	.7631	.8580	.0860	.7676	.8311	.0604	.7307	.8369
EDN[7]	2022	.0449	.7901	.8624	.0723	.7932	**.8415**	.0574	.7208	**.8241**
FocusNet	2024	**.0383**	**.8162**	**.8664**	**.0655**	**.8030**	.8400	**.0531**	**.7302**	.8199
FocusNet+	2024	**.0363**	**.8268**	**.8734**	**.0646**	**.8135**	**.8458**	**.0508**	**.7442**	**.8280**

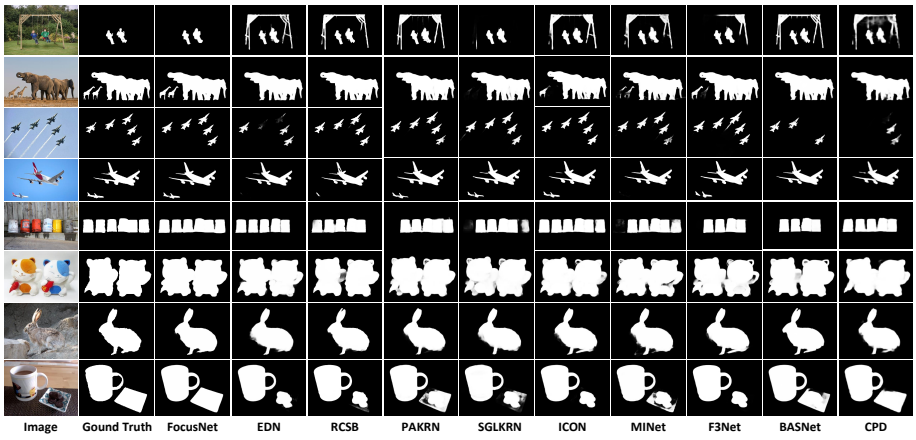

Fig. 7. Visual comparison between our method and other SOTA approaches. Our method demonstrates a clear advantage in scenarios involving multiple targets.

Visual Comparison To further substantiate the efficacy of our method, we offer a series of visual comparisons with other advanced methods. The first image emphatically demonstrates our model's ability to effectively attend to and precisely locate salient targets. The subsequent four images corroborate our model's proficiency in handling multiple targets with similar levels of salience, ensuring that information is not lost on smaller or low-contrast targets. The final two figures demonstrate the superior integrity of our method in segmenting salient objects.

The final trio of images highlights our method's refinement of edges. It is discernible from these images that our approach delineates object boundaries with clarity, eschewing the blurred and ambiguous edges often encountered with prior methods. The representative images selected in previous chapters, along with feature visualization, effectively validated our method's effectiveness. Herein, we reinforce this claim with an expanded array of images that further demonstrate the robustness of our approach.

4.4 Ablation Studies

In this section, a comprehensive ablation study is conducted to systematically evaluate the proposed method's effectiveness. This study meticulously examines three critical aspects: the efficacy of FocusNet's basic components, the FOCUS module's effectiveness and its generalizability across various architectures. This detailed analysis is designed to elucidate how each component contributes to enhancing the model's overall efficacy.

Effectiveness of the Basic Components The effectiveness of each FocusNet component was rigorously tested through a series of experiments. Initially, a U-

Net based on MobileNetv2 was utilized as the baseline. Subsequently, modules were incrementally integrated to assess their individual contributions. Uniformity in experimental settings was meticulously maintained across all models, barring the specific modules under test. The efficacy of each module was rigorously evaluated using the DUTS-TE and HKU-IS datasets, with comprehensive results presented in 2, confirming the substantial performance enhancements afforded by each module.

Table 2. Ablation experiments on basic model components. The upper section details the ablation studies for individual modules, assessing their impact on the overall performance.(The best results are highlighted in bold, and applied in subsequent tables.)

				DUTS-TE			HKU-IS		
Baseline	CFEM	AFE	FOCUS	MAE↓	wF↑	Sm↑	MAE↓	wF↑	Sm↑
√				.0432	.7939	.8548	.0324	.8826	.9018
√	√			.0395	.8074	.8629	.0306	.8894	.9057
√			√	.0409	.8014	.8590	.0313	.8851	.9039
√	√	√		.0383	.8162	.8664	.0298	.8907	.9063
√	√	√	√	**.0363**	**.8268**	**.8734**	**.0282**	**.8969**	**.9106**

Table 3. Diverse ablation experiments on combining backbones and FRM structure. "Low" indicates features from shallower layers, "High" refers to features from deeper layers, and "CA" stands for Channel Attention.

	DUTS-TE				DUTS-TE		
Guide Methods	MAE↓	wF↑	Sm↑	FRM	MAE↓	wF↑	Sm↑
Baseline	.0383	.8162	.8664	Baseline	.0371	.8214	.8701
Horizontal oncatenation	.0381	.8165	.8658	Low+conv(High)	.0365	.8251	.8721
Horizontal concatenation + Location Loss	.0371	.8214	.8701	Low+CA(High)	**.0363**	**.8268**	**.8734**
FOCUS	**.0363**	**.8268**	**.8734**	CA(Low)+CA(High)	.0364	.8259	.8712

Effectiveness and Generalizability of FOCUS The FOCUS module's effectiveness is demonstrated through three distinct information guidance methods, as depicted in 3. Moreover, the FRM's optimal configuration was ascertained from three varied combinations. Left side of 3 illustrates that a mere combination of backbones leads to an increased parameter count with minimal performance enhancement. Conversely, integrating an auxiliary location loss markedly augments performance, as the location network provides horizontal guidance to the secondary. Furthermore, the addition of the FOCUS module enables shallow layers to receive spatial semantic information from deeper layers, thus facilitating spatially-weighted feature integration that substantially bolsters overall performance.

As show in the right of 3, even basic convolution significantly elevates performance. Channel weighting to the deep layer's output marginally enhances efficiency. However, weighting both shallow and deep layers' channels introduces additional parameters and diminishes performance. Consequently, channel weighting was ultimately applied solely to deep-layer information.

On the left side of 4, we have demonstrated the effectiveness of the FOCUS module across five different backbone networks. Further, in the right side of 4, we validate the universality of our module by applying it to five distinct methods: DSS[40], PoolNet[16], PoolRes2Net[16], F3Net[18], and TRACER[23]. After integrating the FOCUS module, there was a noticeable improvement in performance across all models. However, the enhancements observed with large backbone networks were less significant compared to those with lightweight backbone networks, and it will also bring a lot of additional parameter increases.

Table 4. Generalization Experiments of the FOCUS Module. The table provides a comparison of the FOCUS module across various methods and backbones, demonstrating its adaptability in different architectural contexts.

		DUTS-TE					DUTS-TE		
Methods	Backbone	MAE↓	wF↑	Sm↑	Methods	Backbone	MAE↓	wF↑	Sm↑
FocusNet	MobileNetv2	.0383	.8162	.8664	DSS	ResNet50	.0453	.7622	.8292
FocusNet+	MobileNetv2	**.0363**	**.8268**	**.8734**	DSS+	ResNet50	**.0433**	**.7736**	**.8479**
FocusNet	MobileVit-xxs	.0362	.8311	.8817	PoolNet	ResNet50	.0401	.8076	.8831
FocusNet+	MobileVit-xxs	**.0346**	**.8417**	**.8874**	PoolNet+	ResNet50	**.0374**	**.8231**	**.8847**
FocusNet	EfficientNet-B0	.0319	.8495	.8902	PoolRes2Net	Res2Net50	.0374	.8227	.8867
FocusNet+	EfficientNet-B0	**.0307**	**.8556**	**.8918**	PoolRes2Net+	Res2Net50	**.0353**	**.8284**	**.8917**
FocusNet	ResNet50	.0308	.8520	.8929	F3Net	ResNet50	.0355	.8356	.8883
FocusNet+	ResNet50	**.0296**	**.8614**	**.8974**	F3Net+	ResNet50	**.0343**	**.8417**	**.9061**
FocusNet	Swin-Base	.0241	.8905	.9172	TRACER	EfficientNet-b0	.0376	.8271	.8767
FocusNet+	Swin-Base	**.0240**	**.8926**	**.9188**	TRACER+	EfficientNet-b0	**.0353**	**.8368**	**.8819**

5 Conclusion

In this study, we proposed the FOCUS module, which innovatively enhances salient object detection by mimicking the human visual focusing mechanism. This approach proved particularly effective in lightweight networks, significantly improving their precision and efficiency without the substantial increase in computational demands. Additionally, we have introduced an Ascending Feature Enhancement method that effectively improves the model's localization accuracy by altering the direction of feature flow, without significantly increasing the parameter count. Our extensive testing across multiple datasets demonstrated the robustness and scalability of our method, suggesting its potential for widespread application in real-time systems and resource-constrained environments.

References

1. Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., Yang, R.: Salient object detection in the deep learning era: An in-depth survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(6), 3239–3259 (2021)
2. Hong, S., You, T., Kwak, S., Han, B.: Online tracking by learning discriminative saliency map with convolutional neural network. In: International conference on machine learning. pp. 597–606. PMLR (2015)
3. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.M., Feng, J., Zhao, Y., Yan, S.: Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(11), 2314–2320 (2016)
4. Sun, G., Wang, W., Dai, J., Van Gool, L.: Mining cross-image semantics for weakly supervised semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 347–365. Springer (2020)
5. Wang, W., Shen, J., Yang, R., Porikli, F.: Saliency-aware video object segmentation. IEEE transactions on pattern analysis and machine intelligence **40**(1), 20–33 (2017)
6. Gao, Y., Wang, M., Tao, D., Ji, R., Dai, Q.: 3-d object retrieval and recognition with hypergraph analysis. IEEE transactions on image processing **21**(9), 4290–4303 (2012)
7. Wu, Y.H., Liu, Y., Zhang, L., Cheng, M.M., Ren, B.: Edn: Salient object detection via extremely-downsampled network. IEEE Transactions on Image Processing **31**, 3125–3136 (2022)
8. Wang, Y., Wang, R., Fan, X., Wang, T., He, X.: Pixels, regions, and objects: Multiple enhancement for salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10031–10040 (2023)
9. Gao, S.H., Tan, Y.Q., Cheng, M.M., Lu, C., Chen, Y., Yan, S.: Highly efficient salient object detection with 100k parameters. In: European Conference on Computer Vision. pp. 702–721. Springer (2020)
10. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. Pattern recognition **106**, 107404 (2020)
11. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
12. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6848–6856 (2018)
13. Mehta, S., Rastegari, M.: Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178 (2021)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)

16. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3917–3926 (2019)
17. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9413–9422 (2020)
18. Wei, J., Wang, S., Huang, Q.: F^3net: fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI conference on artificial intelligence. pp. 12321–12328 (2020)
19. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8779–8788 (2019)
20. Ke, Y.Y., Tsubono, T.: Recursive contour-saliency blending network for accurate salient object detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2940–2950 (2022)
21. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
22. Chen, Z., Zhou, H., Lai, J., Yang, L., Xie, X.: Contour-aware loss: Boundary-aware learning for salient object segmentation. IEEE Transactions on Image Processing **30**, 431–443 (2020)
23. Lee, M.S., Shin, W., Han, S.W.: Tracer: Extreme attention guided salient object tracing network (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 12993–12994 (2022)
24. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7479–7489 (2019)
25. Wu, Y.H., Liu, Y., Xu, J., Bian, J.W., Gu, Y.C., Cheng, M.M.: Mobilesal: Extremely efficient rgb-d salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 10261–10269 (2021)
26. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
27. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: proceedings of the IEEE/CVF international conference on computer vision. pp. 9197–9206 (2019)
28. Ghiasi, G., Lin, T.Y., Le, Q.V.: Nas-fpn: Learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7036–7045 (2019)
29. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10781–10790 (2020)
30. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
31. Xu, B., Liang, H., Liang, R., Chen, P.: Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 3004–3012 (2021)

32. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5455–5463 (2015)
33. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 136–145 (2017)
34. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1155–1162 (2013)
35. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 280–287 (2014)
36. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3166–3173 (2013)
37. Cheng, M.M., Fan, D.P.: Structure-measure: A new way to evaluate foreground maps. International Journal of Computer Vision **129**, 2622–2638 (2021)
38. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3907–3916 (2019)
39. Zhuge, M., Fan, D.P., Liu, N., Zhang, D., Xu, D., Shao, L.: Salient object detection via integrity learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3738–3752 (2022)
40. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.: Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3203–3212 (2017)