Boosting Broader Receptive Fields for Salient Object Detection

Mingcan Ma*, Changqun Xia*, Chenxi Xie, Xiaowu Chen, *Senior Member, IEEE* and Jia Li, *Senior Member, IEEE*

***Abstract*—Salient Object Detection has boomed in recent years and achieved impressive performance on regular-scale targets. However, existing methods encounter performance bottlenecks in processing objects with scale variation, especially extremely large- or small-scale objects with asymmetric segmentation requirements, since they are inefficient in obtaining more comprehensive receptive fields. With this issue in mind, this paper proposes a framework named BBRF for Boosting Broader Receptive Fields, which includes a Bilateral Extreme Stripping (BES) encoder, a Dynamic Complementary Attention Module (DCAM) and a Switch-Path Decoder (SPD) with a new boosting loss under the guidance of Loop Compensation Strategy (LCS). Specifically, we rethink the characteristics of the bilateral networks, and construct a BES encoder that separates semantics and details in an extreme way so as to get the broader receptive fields and obtain the ability to perceive extreme large- or small-scale objects. Then, the bilateral features generated by the proposed BES encoder can be dynamically filtered by the newly proposed DCAM. This module interactively provides spacial-wise and channel-wise dynamic attention weights for the semantic and detail branches of our BES encoder. Furthermore, we subsequently propose a Loop Compensation Strategy to *boost* the scale-specific features of multiple decision paths in SPD. These decision paths form a feature loop chain, which creates mutually compensating features under the supervision of boosting loss. Experiments on five benchmark datasets demonstrate that the proposed BBRF has a great advantage to cope with scale variation and can reduce the Mean Absolute Error over 20% compared with the state-of-the-art methods.**

***Index Terms*—Salient object detection, receptive field, bilateral extreme stripping, loop compensation.**

I. Introduction

In recent years, methods based on deep learning [1]–[6] have made great progress in the Salient Object Detection (SOD) field by virtue of powerful feature representation. For example, methods [7]–[13] based on Convolutional Neural Network (CNN) usually utilize multi-layer convolutions to extract global semantics and local detailed features at the same time. Methods [14] based on vision transformers abandon the current vision attention mechanism and provide a way to understand images from a holistic perspective. However, existing methods still encounter the challenge: It is difficult to balance the segmentation effects of objects with scale variation. As shown in Fig. 1, CNN-based methods such as PA-KRN [8] may generate more failure cases when dealing with large-scale objects, while transformer-based methods such as VST [14] may have trouble handling small-scale objects. Specifically, the predicted maps with large-scale objects have better F-measure but higher Mean Absolute Error (MAE), while those with small-scale objects have the opposite performance. That is, there are *asymmetric segmentation requirements* for extremely large or small objects.

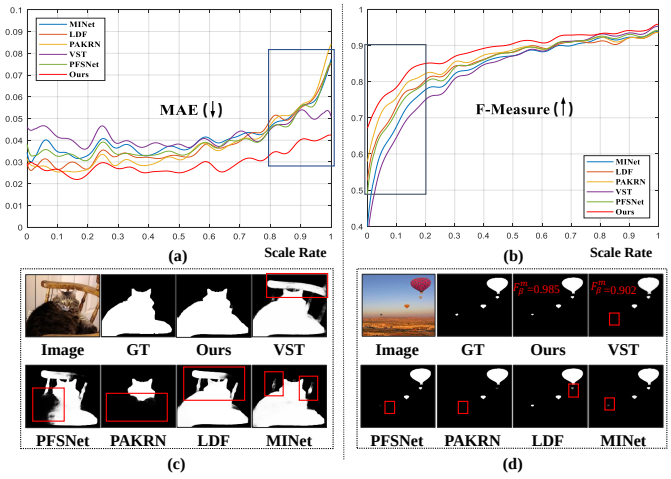

Fig. 1. Comparisons between our BBRF and six state-of-the-art methods in the scenarios with scale variation. Scale Rate represents relative value of foreground pixel area ratio. (a) and (b): Quantitative comparison under MAE and mean F-measure. (c) and (d): Visual comparison on large- and small-scale objects.

To solve this challenge, the state-of-the-art methods seek to design multi-scale modules or models. For instance, PoolNet [15] proposes a multi-branch module to extract multi-scale features by compressing features to different sizes through pooling operations with different sampling rates. PFANet [16] extracts rich-scale information in a multi-decoder model through decoupling semantics and details to different decoders. MINet [10] and PFSNet [7] make full use of the relationship between adjacent features to obtain scale-aware information while avoiding the introduction of noise. UTA [17] proposes a Gated Multi-Scale (GMS) module to extract multi-scale information separately in an efficient way. VST [14] utilizes a

Mingcan Ma and Chenxi Xie are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, 100191, China.

Changqun Xia is with Peng Cheng Laboratory, Shenzhen, 518000, China.

Xiaowu Chen and Jia Li are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, 100191, China, and also with Peng Cheng Laboratory, Shenzhen, 518000, China.

*Mingcan Ma and Changqun Xia contributed equally to this work. Corresponding author: Jia Li (E-mail: jiali@buaa.edu.cn).

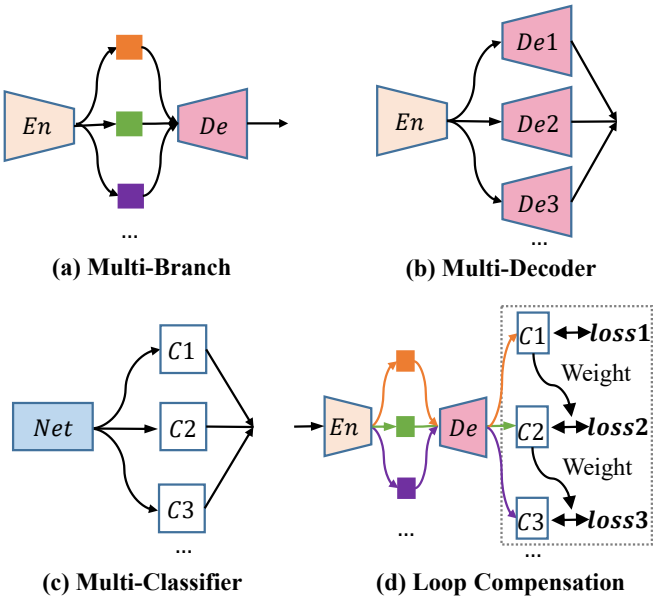

Fig. 2. Comparisons between our Loop Compensation Strategy and other related methods. (a): FAM of PoolNet [15]. (b): Decoder of PFANet [16]. (c): GMS proposed in UTA [17]. (d): Proposed Loop Compensation Strategy. Lines with different colors indicate different decision paths selected by a switch.

pure transformer architecture to extract more powerful features from a holistic perspective. In spite of impressive performance, these methods may suffer from the following weaknesses:

1) The local perception mechanism of CNN methods limits the range of their receptive fields, while the attention mechanism between spatial regions of pure transformer is difficult to ensure local details and efficiency. This contradiction makes it difficult to efficiently segment objects with varying scales. They may have difficulty balancing global perspective, precise detail, and model efficiency.

2) As shown in Fig. 2, representative methods set up different forms of multi-stream structures to generate expressive features. However, they often set one unique decision path or ignore the correlation of different decision results. Here, the decision path refers to a structure path of the network that can generate results independently. Besides, the undifferentiated training process between different paths or decoders such as Fig. 2 (a) and (b) limits the scale specificity of the models.

Therefore, Constructing receptive fields with elasticity and scale specificity may be of great significance for comprehensive segmentation of objects with scale variation. For more comprehensive receptive fields, this paper proposes a framework named BBRF for Boosing Broader Receptive Fields, which includes a Bilateral Extreme Stripping encoder with Dynamic Complementary Attention Modules and a Switch-Path Decoder with a Loop Compensation Strategy.

First, we rethink how to utilize the advantages of bilateral structures for dealing with scale variation and propose Bilateral Extreme Stripping to build a distinctive bilateral network based on CNNs and transformers for the broader receptive fields. That is, we separate semantic and detail information as much as possible. The semantic branch can almost ignore high-resolution details, while the detail branch only efficiently extracts shallow high-resolution features. In this way, deep semantics and shallow details can be efficiently extracted in the transformer and CNN parts, respectively. To accommodate performance and efficiency, we choose a lightweight CNN and a simplified transformer to construct our BBRF. The lightweight CNN keeps a high-resolution input unchanged to extract local details, while the simplified transformer branch is mainly responsible for extracting global semantics among regions on the premise of the lowest possible input resolution. In our setting, the input resolution of detail branches can be set to a maximum of seven times the semantic branch. In this way, we can better take into account the efficiency and more comprehensive receptive fields.

Second, the Dynamic Complementary Attention Module is proposed to provide dynamic attention enhancement between our bilateral branches of Bilateral Extreme Stripping encoder and realize more effective filtering of the broader bilateral features. Unlike other self-attention modules, DCAM emphasizes dynamic associations between complementary branches. The semantic branch provides dynamically adaptive spatial-level attention weights for the detail branch, and the detail branch simulates fluctuations between channels for the semantic branch. Our DCAM emphasizes the spatial attention of the CNN and the channel flexibility of the transformer by weighting them mutually, which can provide more elastic receptive fields. With DCAM, our BES encoder can not only compensate for the limitations of convolutional receptive fields but also build dynamic associations between transformer feature channels.

Third, we further propose a novel and effective loop compensation strategy to boost scale-specific views based on the broader receptive fields. Different from Multi-Branch and Multi-Decoder modules in Fig. 2 (a) and (b), LCS adopts Loop chain compensation manner, that is, LCS serially trains one unique decision path in each training iteration. Unlike GMS [17], our LCS assigns convolutions with different dilated rates for each path to enhance the ability of feature representation. More importantly, when calculating the loss for each path, our LCS focuses on the pixel regions where previous decision paths were incorrectly predicted. Accordingly, chain relations between adjacent decision paths can be constructed. All decision paths can form a sequential chain loop, which can generate scale-specific receptive fields. With almost equivalent parameters, LCS can improve the performance of the proposed BBRF by a large margin.

The experimental results validate the performance and efficiency of our BBRF, especially in handling scale variations. Here, we highlight the following contributions.

1) We innovatively analyze the *asymmetric segmentation requirements* of extremely large- and small-scale objects, and consider more comprehensible receptive fields from both scale elasticity and scale specificity enhancement. 2) We rethink the advantages of bilateral structures and construct a Bilateral Extreme Stripping network equipped with the proposed Dynamic Complementary Attention Modules for more elastic receptive fields including extremely large- or small-scale views. 3) We further propose a novel yet effective Loop Compensation Strategy to boost scale-specific views based on the broader

receptive fields, which utilizes boosting loss to make each decision path in SPD pay more attention to the predicted errors of the previous path through boosting loss and break the synchronization relationship of the error back-propagation process. 4) Our method can better handle scale variation and achieve a significant performance improvement over 16 state-of-the-art methods. In particular, the MAE metric of our method reaches **0.022** on ECSSD and **0.025** on DUT-TE, which is **29%** and **26%** lower than the best results before.

II. Related work

Salient Object Detection aims to segment the visually significant objects in the image, which can contribute to related downstream tasks [18]–[20]. Despite the landmark progress of salient object segmentation methods [21]–[27] based on deep learning, scale variation has been a worthy challenge in the field of SOD. Multi-scale object perception [16], [28]–[32] requires simultaneous extraction of global semantics and local details. To balance semantics and details, existing methods often integrate multi-level features or design multi-scale modules. We will review related work from these two aspects.

A. Multi-feature integration

Recently, methods based on multiple features integration have achieved impressive progress in SOD and related fields [31], [33]–[39]. For example, DSNet [40] proposes to dynamically filter RGB-D image features using global guidance information to obtain a more complete feature representation. RRNet [41] proposes a multi-scale attention mechanism to integrate semantic relational features. GLNet [42] considers co-aggregating features from multiple images for Co-Salient Object Detection. F3Net [43] proposes Fusion, Feedback and Focus to detect salient objects and has achieved the best performance at the time. ITSDNet [44] proposes to construct the association between salient features and contour features, which can effectively optimize the boundaries of salient objects. LDF [45] strips the features based on the distance from the edge and iteratively optimizes the predicted maps for more accurate results. PA-KRN [8] proposes to initially locate the salient objects, and then segment them carefully. This strategy can better balance detailed features and semantic features.

Besides, VST [14] utilizes the transformer [46] to detect salient objects for the first time, which can construct patch-level global features. In this way, the network can obtain a global perspective. It can be seen that CNN methods pay more attention to boundaries and details, while methods based on transformers pay more attention to semantic features. In contrast, we explore the differences between CNNs and transformers in terms of semantics, details and computational complexity, so as to make a better trade-off among them.

In addition, Conformer [47] first proposes to construct a dual branch network based on transformer and CNN. However, this paper rethinks such bilateral structures with the following essential differences: 1) Efficient extreme stripping design: the detailed feature resolution is seven times than the semantic resolution in our framework, while the input resolution of bilateral branches of the Conformer [47] does not differ much, which may lead to redundant information and affect efficiency. 2) Independent stripping branches: compared to Conformer, which interpolates features in bilateral branches, our independent features extraction of extreme-scale objects is more suitable for their asymmetric segmentation requirements. 3) Dynamic complementary attention mechanism: compared with the way that the converter directly converts bilateral features, we use the proposed DCAMs to filter features to better accommodate the scale variation of random samples.

B. Multi-scale perception

In the SOD field, enhancing multi-scale perception capabilities by constructing multi-branch modules is an effective strategy to improve feature representation capabilities. For example, Pyramid Pooling Module (PPM) [48] is a typical multi-branch module, in which each branch is down-sampled to a different resolution to achieve multi-scale feature representation. Atrous Spatial Pyramid Pooling (ASPP) [49] module utilizes convolutions with different dilated rates in each branch to achieve multi-scale feature extraction. PoolNet [15] uses PPM for SOD for the first time and achieves the desired performance results. BANet [1] proposes to improve ASPP and obtains more powerful feature representation. DINet [50] proposes to use parallel dilated convolutions to enrich the network receptive field. PSGLoss [51] proposes to adaptively capture multi-scale features though a branch-wise attention mechanism. MMNet [52] explores a multi-stage fusion strategy to obtain more complete multi-scale features.

Although the modules proposed by the above methods can effectively increase the feature expression ability and achieve impressive segmentation results, they suffer from the performance bottleneck since they ignore to balance the asymmetric requirements of object segmentation at different scales. The original intention of the multi-branch model is to allow different branches to capture specific features, but the previous methods seldom consider the independence of each branch and the mutual relationship between different branches. Besides, although they can effectively expand the receptive fields, it may be difficult for them to break the limitation of local attention mechanism. In contrast, we propose a loop compensation strategy to enhance the complementarity between different paths in decoder. Each path calculates the error separately and weights the error area of the previous decision path.

III. METHODOLOGY

A. Overview

In this section, we will introduce the overview of the framework. Our BBRF is built based on the encoder-decoder structure but has its own uniqueness. First, we construct a BES encoder for feature extraction and four DCAMs for feature fusion. Distinguishing from existing bilateral models, our BES encoder considers the extreme stripping of semantic and detailed features, which stems from the following observations: 1. The emergence of the vision transformers provides a new way for global feature perception, but the construction

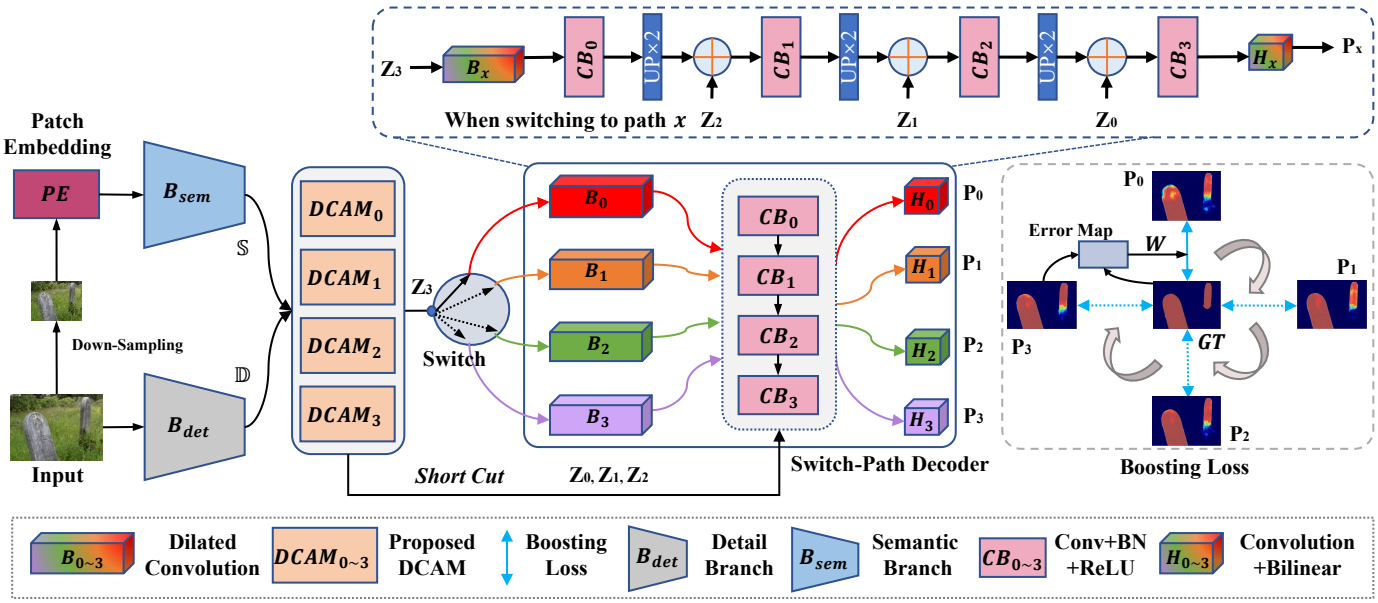

Fig. 3. The pipeline of the proposed method. This figure shows the training process of our BBRF. Different colors in the Switch-Path Decoder indicate different paths.The paths are trained one by one, and the error area of the previous path is weighted in each iteration.

of global semantics leads to a large model computation. Therefore, it is difficult to use the existing transformer models directly for efficient SOD tasks such as Sal100K [53]. 2. There are several contradictory points in semantic and detailed information. Deeper global iteration tends to enhance the semantic richness while polishing at shallow high resolution is more suitable for detail extraction. In brief, by separating details and semantic in an extreme manner, our BES encoder and DCAM can efficiently obtain more elastic and broader receptive fields.

The Switch-Path Decoder shown in Fig. 3 represents the unique structure of our framework. Unlike existing typical decoders, SPD contains multiple decision paths. The design idea of SPD is based on our considerations of existing multi-branch structures. These structures are similar to existing integration models, and the key to improve the overall performance is to exploit the uniqueness of multi-branches and reduce redundancy. Existing multi-branch structures rarely consider the prediction relationships between different branches. Different branches tend to receive the same action of the gradient back-propagation process. This approach greatly limits the independence of different branches. For this reason, we designed SPD and Loop Compensation Strategy to explicitly constrain the characteristics of different branches.

The forward inference process of the framework is shown in Fig. 3. For a given input $I \in \mathbb{R}^{H\times W\times 3}$, we sequentially perform down-sampling and patch embedding operations [54] to get I_p and then input I and I_p into B_{det} and B_{sem}, respectively. B_{det} can get the intermediate feature set $\mathbb{D} = \{\mathbf{D}_i | i = 0, 1, ..., K-1\}$ of the corresponding blocks. Similarly, the feature set of B_{sem} can be obtained: $\mathbb{S} = \{\mathbf{S}_i | i = 0, 1, ..., K-1\}$. We refer to the design of the number of blocks in ResNet [55] and set K to four. Since the feature shapes of B_{det} and B_{sem} are different, the features of B_{det} need to be converted, which can be expressed as $\mathbf{S}_i^{'} = \mathcal{C}(\mathcal{U}(\mathcal{R}(\mathbf{S}_i)))$, where $\mathcal{R}$ represents the reshaping of vector group features into planar features, $\mathcal{U}$ represents up-sampling operation, and $\mathcal{C}$ represents 1×1 convolution. Then feature sets $\mathbb{S}^{'}$ and $\mathbb{D}$ are filtered through DCAM to generate feature set $\mathbb{Z}$, where $\mathbf{Z}_i = DCAM_i(\mathbf{S}_i, \mathbf{D}_i), i = 0, 1, ..., K-1$. The deepest feature $\mathbf{Z}_3 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C}$ only passes through one decoding path, and other paths will not calculate the gradient here. Assuming that the current x-th path is activated, the error-weighted back-propagation will be performed on the result according to the error situation of previous decision path. Let $\mathbf{B}_x$ denotes the feature generated by B_x of Fig. 3. $\mathbf{P}_x$ represents the predicted map corresponding to H_x of Fig. 3. Then $\mathbf{P}_x = decoder(\mathbf{Z}, \mathbf{B}_x)$, where $decoder$ composed of Convolution Blocks represents one of the decision path as show in Fig. 3.

The testing process is different from the training process, and the gradient does not need to be saved. The final result can be generated by the prediction set $\mathbb{P}$, which can be expressed as $\mathbf{F}_{out} = \sigma(\sum_{i=0}^{3} \mathbf{P}_i)$, where σ denotes Sigmoid activation function. The testing process will activate all decision paths.

B. Bilateral Extreme Stripping (BES)

Based on the observation of the phenomenon in Fig. 1, the existing SOD methods may produce more failure cases when dealing with extremely large- or small-scale objects. Moreover, the segmentation process has an asymmetric segmentation requirements in the extreme-scale range. Therefore, we propose a Bilateral Extreme Stripping encoder based on existing bilateral networks [56] for the broader receptive fields. This framework partitions the segmentation difficulties at the extreme-scale range into different branches. Our BES encoder contains a semantic branch and a detail branch. The semantic branch makes full use of the global view of the transformer model to extract global semantic features and enhance the

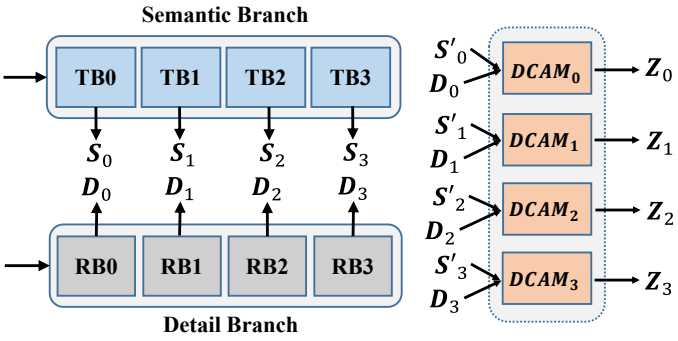

Fig. 4. Feature filtering process of our BES encoder. The corresponding features are consistent with Fig. 3.

receptive fields of the overall network. The detail branch is responsible for extracting local details. The advantage of our BES is to strip semantic information and details. Extracting detailed information often requires a higher input resolution, while obtaining global semantics can appropriately reduce the resolution. In this way, we can achieve a better trade-off between details, semantics, and efficiency. Moreover, our BES can apply the existing complex backbone to more downstream tasks in a more efficient form.

Specifically, Fig. 4 shows the detailed structure of our BES encoder, and the symbols used are consistent with the Fig. 3. It can be seen that our BES merges the features at the same level of the two branches. We choose ResNet-18 [55] to build the detail branch. In order to enrich the detailed information extracted by this branch, we remove the first down-sampling operation to obtain more effective features at high resolution. In this way, the resolution of the features output by i-th layer is $\frac{1}{2^i}$ of the initial resolution. As for the semantic branch, we utilize Swin Transformer [54] to extract global features. Since we only expect the semantic branch to provide more global semantic associations, this allows us to choose a suitably small input resolution for that branch. Thus, the input resolution of the semantic branch is set to 56×56, and then passes through the self-attention modules of the transformer in turn.

Notably, Conformer [47] utilizes a similar idea of bilateral networks [56], [57] to combine CNNs and transformers, which has achieved state-of-the-art performance in visual recognition tasks. But our method is designed for *asymmetric segmentation requirements* and has the following specific considerations: 1) Although the complementary features of these two types of models can be integrated, the Conformer does not make full use of the advantages of bilateral networks to accelerate calculations. It is a luxury to extract semantics and details separately with the same initial resolution input, especially transformers are not sensitive to details compared to CNNs. 2) To cope with the *asymmetric segmentation requirements*, we maintain the exclusivity of functions between the internal bilateral branching structures. 3) In contrast to direct bridging of bilateral features in Conformer, we propose a dynamic attention filtering method for complementary features.

C. Dynamic Complementary Attention Module (DCAM)

Our Bilateral Extreme Stripping encoder can efficiently produce both low-resolution semantic features and high-resolution detailed features. However, this introduces another problem: how to combine the bilateral features to obtain more elastic receptive fields. We propose DCAM to solve this problem. Unlike other feature fusion modules, our DCAM is designed to address the different characteristics of CNNs and transformers as well as the difference in resolution of semantics and details. We observe that global attention of the transformer calculates the correlation between all patches in the spatial level through the vector inner product, while the attention mechanism of CNN establishes the connection between all channels in the local space. The former can hardly express the weight proportion between channels, while the latter is difficult to calculate the weight coefficient of the whole spatial range. Therefore, CNN features are used to generate dynamic channel-wise weights for the transformer to make up for the lack of correlation among channels. Transformer features are utilized to obtain spatial weights for CNN to supplement the spatial perspective. The dynamic attention mechanism can adapt to the features of different models while bridging the resolution differences of the bilateral features.

As shown in Fig. 5, it is assumed that $\mathbf{X} \in \mathbb{S}^{'}$ and $\mathbf{Y} \in \mathbb{D}$ are features to be fused. We first use $\mathbf{X}$ to generate the spatial weights for $\mathbf{Y}$, which can be expressed as $\mathbf{X}^{'} = \sigma(\mathcal{U}(\mathcal{C}(\mathbf{X})))$. Here $\mathcal{U}$ represents the up-sampling operation, $\mathcal{C}$ represents 1×1 convolution, and σ represents the Sigmoid activation function. Then we generate the channel weights of feature $\mathbf{X}$, which can be formulated as $\mathbf{Y}^{'} = \sigma(\mathcal{G}(\mathcal{C}(\mathbf{Y})))$. $\mathcal{G}$ denotes global average pooling operation. Next, DCAM will utilize dynamic weights to enhance bilateral features, which can help our BES encoder form a more elastic feature representation. The specific process can be formulated as follows:

$$\mathbf{M} = \mathcal{C}_{br}(\mathcal{U}(\mathbf{Y}^{'} \otimes \mathbf{X}) \copyright (\mathbf{X}^{'} \otimes \mathbf{Y})), \tag{1}$$

where $\mathcal{C}_{br}$ represents a convolution with a Batch Normalization and a ReLU activation. $\otimes$ represents pixel-wise multiplication. $\copyright$ denotes feature concatenation. Finally, the output feature $\mathbf{Z}$ is generated through the residual structure, which can be represented as:

$$\mathbf{Z} = \mathcal{C}_{br}(\mathbf{M} \oplus \mathcal{U}(\mathcal{C}(\mathbf{X})) \oplus \mathcal{C}(\mathbf{Y})), \tag{2}$$

where $\oplus$ denotes pixel-wise addition. Since the features obtained by transformer $\mathbf{X}$ contain rich spacial semantics and CNN-based features $\mathbf{Y}$ are more capable of constructing channel associations, we use the spatial semantics of $\mathbf{X}$ to select the spatial information of $\mathbf{Y}$, and use the channel-wise vector generated by $\mathbf{Y}$ to adjust the channel weight of $\mathbf{X}$. The final result is obtained by fusing all the information through the residual structure.

To demonstrate the effectiveness of DCAM, we visualize the features before and after processing. Fig. 5 shows the features visualization before and after DCAM. $\mathbf{X}$-Low and $\mathbf{X}$-High indicate the situation where $\mathbf{X}$ takes the first layer feature and the deepest layer feature, respectively. It can be seen that the BES features have been severely differentiated, but the

TABLE I

QUANTITATIVE COMPARISON TABLE WITH THE LATEST METHODS ON MULTIPLE INDICATORS, INCLUDING THE MAX AND MEAN F-MEASURE (F_β^* AND F_β^m THE LARGER THE BETTER), MAE (THE SMALLER THE BETTER), E-MEASURE (E_ξ, THE LARGER THE BETTER), AND S-MEASURE (S_m, THE LARGER THE BETTER) . THE BEST AND SECOND BEST RESULTS ARE MARKED IN RED, AND BLUE, RESPECTIVELY.

Method	ECSSD (1000)					HKU-IS (4447)					DUTS-TE (5019)					DUT-OMRON (5168)					PASCAL-S (850)				
	$F_\beta^*\uparrow$	$F_\beta^m\uparrow$	MAE $\downarrow$	$E_\xi\uparrow$	$S_m\uparrow$	$F_\beta^*\uparrow$	$F_\beta^m\uparrow$	MAE $\downarrow$	$E_\xi\uparrow$	$S_m\uparrow$	$F_\beta^*\uparrow$	$F_\beta^m\uparrow$	MAE $\downarrow$	$E_\xi\uparrow$	$S_m\uparrow$	$F_\beta^*\uparrow$	$F_\beta^m\uparrow$	MAE $\downarrow$	$E_\xi\uparrow$	$S_m\uparrow$	$F_\beta^*\uparrow$	$F_\beta^m\uparrow$	MAE $\downarrow$	$E_\xi\uparrow$	$S_m\uparrow$
BASNet [2]	.942	.880	.037	.921	.916	.928	.895	.032	.946	.909	.860	.791	.048	.884	.866	.805	.756	.056	.869	.836	.857	.775	.076	.847	.832
PoolNet [15]	.944	.914	.039	.924	.922	.933	.896	.032	.949	.910	.880	.811	.040	.889	.878	.808	.746	.056	.863	.828	.869	.823	.074	.850	.847
CPD [3]	.939	.917	.037	.925	.918	.925	.891	.034	.944	.905	.865	.805	.043	.887	.869	.797	.747	.056	.866	.825	.864	.824	.072	.849	.842
BANet [1]	.945	.880	.035	.928	.916	.931	.895	.032	.950	.909	.872	.791	.040	.892	.866	.803	.756	.059	.860	.836	.870	.775	.070	.855	.832
GateNet [58]	.945	.916	.040	.924	.920	.933	.899	.033	.949	.915	.888	.807	.040	.889	.885	.818	.746	.055	.862	.838	.875	.825	.068	.852	.852
U2Net [59]	.951	.892	.033	.924	.928	.935	.896	.031	.948	.916	.873	.792	.045	.886	.874	.823	.761	.054	.871	.847	.862	.772	.076	.841	.836
DFI [60]	.949	.920	.035	.924	.927	.934	.902	.031	.951	.920	.886	.814	.039	.892	.887	.818	.752	.055	.865	.839	.885	.837	.065	.857	.861
MINet [10]	.947	.924	.033	.927	.925	.935	.909	.029	.953	.919	.884	.828	.037	.898	.884	.810	.755	.055	.865	.833	.865	.835	.064	.852	.851
GCPANet [61]	.948	.919	.035	.920	.927	.938	.898	.031	.949	.920	.888	.817	.038	.891	.891	.812	.748	.056	.860	.839	.876	.833	.061	.850	.861
ITSDNet [44]	.947	.895	.034	.927	.925	.934	.899	.031	.952	.917	.883	.804	.041	.895	.885	.821	.756	.061	.863	.840	.876	.792	.064	.853	.856
LDF [45]	.950	.930	.034	.925	.924	.939	.914	.027	.954	.919	.898	.855	.034	.910	.892	.820	.773	.051	.873	.838	.874	.843	.059	.865	.856
PFSNet [7]	.952	.932	.031	.928	.930	.943	.919	.026	.956	.924	.896	.847	.036	.902	.892	.823	.774	.055	.875	.842	.875	.837	.063	.856	.854
PSGLoss [51]	.949	.932	.031	.928	.925	.938	.918	.027	.958	.919	.886	.849	.036	.908	.883	.811	.771	.052	.870	.831	.879	.848	.061	.858	.856
PurNet [62]	.945	.921	.035	.925	.925	.936	.904	.030	.950	.917	.878	.823	.039	.897	.880	.814	.756	.051	.868	.841	.873	.827	.068	.851	.843
VST [14]	.951	.920	.033	.918	.932	.942	.900	.029	.953	.928	.890	.818	.037	.892	.896	.825	.756	.058	.861	.850	.875	.829	.061	.837	.865
PA-KRN [8]	.953	.931	.032	.924	.928	.943	.920	.027	.955	.923	.907	.865	.033	.916	.900	.834	.793	.050	.885	.853	.873	.838	.066	.857	.852
BBRF-tiny	.958	.948	.024	.932	.935	.947	.936	.023	.962	.927	.910	.890	.026	.926	.900	.840	.810	.039	.885	.847	.890	.865	.051	.864	.866
BBRF	.963	.950	.022	.934	.939	.958	.945	.020	.965	.935	.916	.893	.025	.927	.908	.843	.814	.042	.887	.855	.891	.869	.049	.867	.871

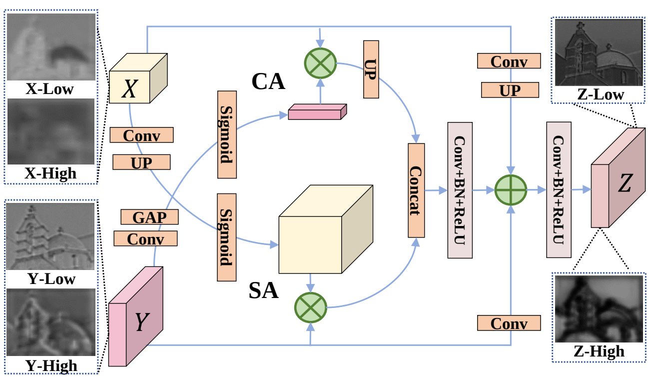

Fig. 5. Illustration of the Cross-Domain Attention Fusion Module. **X** and **Y** represent the output features of the transformer and CNN, respectively. The feature maps denote the fusion effect when **X** and **Y** take low-level or high-level features.

semantics and details are present in features X and Y. However, after DCAM processing, the feature maps can represent the salient objects more clearly and accurately.

D. Loop Compensation Strategy (LCS)

Bilateral Extreme Stripping and Dynamic Complementary Attention Module can adaptively filter extreme-scale features and generate broader receptive fields. To optimize the segmentation effect, we further propose LCS to enhance the perception of different scale ranges and suppress error transmission. On the one hand, We divide features of different scale ranges into different decision paths through setting convolutions with different dilated rates. On the other hand, we adopt the random training process to suppress error transmission between adjacent decision paths. With our LCS, we hope that each decision path can acquire scale-specific features and different paths could complement each other.

Compared with existing multi-branch modules or models, the decision path of LCS is more flexible. Different decision paths, although contributing most of the parameters, have parameter-independent dilated convolution and prediction heads. The training process of different paths is randomly and independently performed. The previous methods seldom explore ways of partial network training. Each decision path in the proposed Switch-Path Decoder can be trained separately. This can make full use of the differences in training data and different loss functions to obtain path specificity. In other words, each path of SPD is no longer constrained by the identical training data and error gradient. Meanwhile, the parameter sharing, especially our SPD, can effectively ensure the efficiency and calculation speed of the framework. LCS can be explained in two parts: Switch-Path Decoder and Boosting Loss. We will introduce these two parts in detail.

Switch-Path Decoder aims to strip multi-branch modules trained in parallel into multiple decision paths, thereby reducing the gradient correlation and realizing the boosting operation between paths. Boosting loss is designed for the integration of multiple decision paths. Its goal is to enhance the complementary relationship of different paths and enhance the overall effect of the ensemble model. As shown in Fig. 2, we set a prediction head for each decision path, so that each path can be trained separately. After path stripping, the prediction process becomes a multi-path voting mechanism. During training, the complementarity among paths can be enhanced by Boosting Loss. When testing, the final saliency map can be generated from the voting results of multiple paths. Only by introducing affordable parameters and calculations,

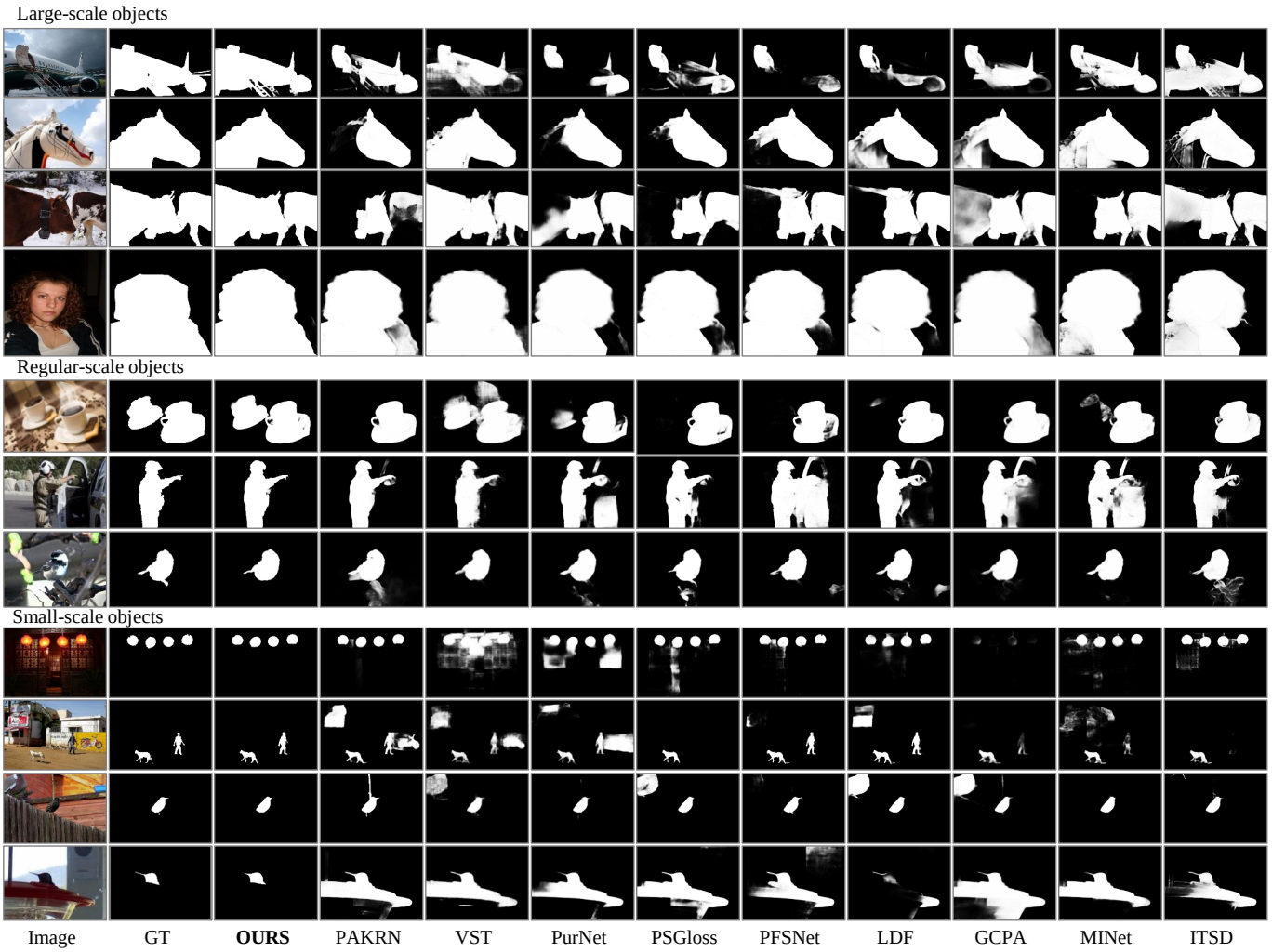

Fig. 6. Visual comparison between the proposed method and the state-of-the-art methods. Our method can adapt better to input samples of different scale ranges. Moreover, our BBRF can more precisely localize small-scale objects in multi-objects scenes, while being able to accurately segment the details of large-scale objects.

LCS can realize the division of independent decision paths. Each decision path has proprietary dilated convolutions and prediction heads. As shown in Fig. 3, SPD specifies a decision path through a switch. Any selected decision path is the structure shown in the top of Fig. 3. Then we let $\mathbf{Z}_4$ go through a specific dilated convolution to get the feature $\mathbf{B}_x$. We merge the residual feature set $\mathbb{Z}$ and $\mathbf{B}_x$ like FPN and obtain the filtered features. Finally, we can get predicted maps after H_x. Here, $x \in \{0, 1, 2, ..., N-1\}$. In addition, the simplification of a single path can speed up the convergence of the model.

Boosting Loss aims to enhance the complementarity between adjacent paths. During each iteration, we will randomly select a path to save the gradient for training, and predict the prediction result of the previous path to calculate the error weight. Each iteration strengthens the chain compensation relationship between adjacent paths of the model, and finally achieves loop compensation of multiple decision paths, so as to obtain more robust feature presentation. As shown in Fig. 3, assume that $x \in \{0, 1, ..., N-1\}$ represents the path number that needs to be trained in the current iteration. Then this iteration will only perform error back-propagation on the x-th path. And this path will consider the prediction error of the previous path as a weight $\mathbf{w} \in \mathbb{R}^{H \times W}$. Here, H and W indicate the height and width of the image. In this way, we can highlight the pixel positions that were previously predicted incorrectly. The calculation process of $\mathbf{w}$ can be denoted as:

$$\mathbf{w}_x = \mathcal{B}(\mathbf{P}_{(x-1+N)\%N}, g) + 1, \tag{3}$$

where $g \in \mathbb{R}^{H \times W}$ denotes the ground-truth map. $\mathcal{B}$ denotes the pixel-level Binary Cross Entropy [63], which can be denoted as:

$$\mathcal{B}(p, g) = -(g \otimes log(p) \oplus (1-g) \otimes log(1-p)), \tag{4}$$

where $p \in \mathbb{R}^{H \times W}$ denotes predicted map. log represents a pixel-level logarithmic operation. The weight $\mathbf{w}_x$ records the weaknesses of the x-th decision path, and these weaknesses will be strengthened during the training process of the adjacent paths. The path compensation effect of our LCS is reflected in the weighting process of the same loss across multiple paths.

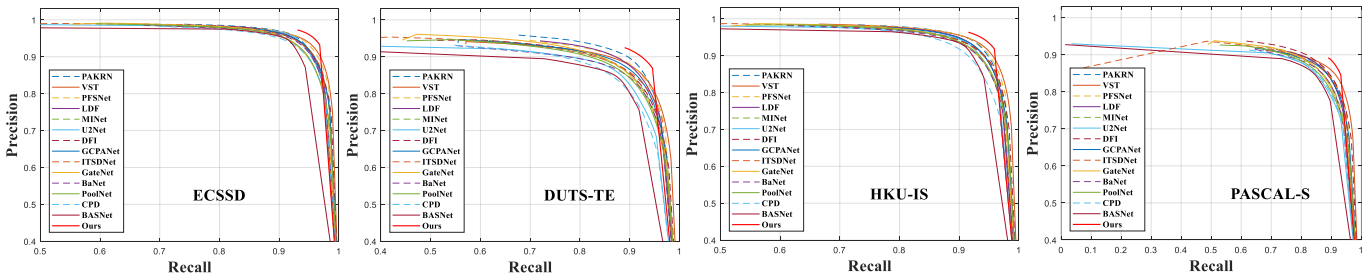

Fig. 7. Comparison of PR curves between our method and the state-of-the-art methods under four benchmark dataset. The figure shows that the proposed method can better balance accuracy and recall.

The Boosting Loss can be calculated according to the weight $\mathbf{w}_x$. It can be expressed as:

$$\mathcal{L}_b(p, g, w) = \mathcal{L}_{wbce}(p, g, \mathbf{w}_x) + \mathcal{L}_{wiou}(p, g, \mathbf{w}_x), \quad (5)$$

where $\mathcal{L}_{wbce}$ and $\mathcal{L}_{wiou}$ represent weighted Binary Cross Entropy loss and weighted Intersection over Union loss, which has been widely used in many methods such as [7], [17], [43], [45]. They can be expressed as:

$$\mathcal{L}_{wbce}(p, g, \mathbf{w}_x) = \frac{\mathcal{S}(\mathcal{B}(p, g) \otimes \mathbf{w}_x)}{\mathcal{S}(\mathbf{w}_x)}, \quad (6)$$

$$\mathcal{L}_{wiou}(p, g, \mathbf{w}_x) = 1 - \frac{\mathcal{S}((p \otimes g) \otimes \mathbf{w}_x)}{\mathcal{S}(((p \oplus g) - (p \otimes g)) \otimes \mathbf{w}_x)}, \quad (7)$$

where $\mathcal{S}$ represents the operation of summing all pixels. Other symbols are consistent with the previous description.

E. Learning objective

We utilize the sum of Binary Cross Entropy (BCE) [63] and Intersection over Union (IoU) [64] as the loss function, which is widely used in LDF [45], etc. In addition to the final prediction map, we will also supervise the output features of each Convolution Block (CB) in Fig. 3. The total loss function can be expressed as:

$$\mathcal{L}_{tot} = \sum_{i=1}^{M} \mathcal{L}_b(F_i, g, \mathbf{w}_x) + \mathcal{L}_b(P_x, g, \mathbf{w}_x), \quad (8)$$

where $F_i \in \mathbb{R}^{H \times W}$ represents the prediction map of each CB module, and $P_x \in \mathbb{R}^{H \times W}$ represents the predicted map of path x. M denotes the number of CB modules.

IV. Experiment

A. Experiment setting

Datasets and metrics. This paper involves the following datasets: DUT-OMRON [65] with 5,168 images, ECSSD [66] with 1,000 images, HKU-IS [67] with 4,447 images, PASCAL-S [68] with 850 images, DUTS-TE [69] with 5,019 images, DUTS-TR [69] with 10,553 images. Consistent with the latest methods [45], [70], we choose DUTS-TR for training and other datasets for verification.

Six widely used evaluation metrics are selected to evaluate the performance of our method and existing state-of-the-art methods. The first metric is the precision-recall curve. We use the precision-recall curve to evaluate the prediction results comprehensively. The second is MAE [71], which is defined as the pixel-wise average absolute error between predicted maps and ground-truth maps. To comprehensively consider recall and precision, we use the max and mean F-measure (F_β^*, F_β^m) [65] to emphasize the proportional relationship between recall and precision. Besides, we also utilize maximum Enhanced-alignment Measure (E_ξ) [72] and Structure Measure (S_m) for more comprehensive comparisons.

Implementation details. We use an NVDIA GTX 2080Ti GPU to train our network. By selecting different input resolutions, we design two models with different calculation costs. The input resolutions of BBRF and BBRF-tiny in the detail branch are 352^2 and 224^2 respectively. The resolutions in the semantic branch are both 56^2. The maximum learning rate of the backbone is 0.004, and the other parts are expanded by ten times. We use Warm-up and linear decay strategies to adjust learning rate [45]. The optimization method uses Stochastic Gradient Descent. Batch size is set to 26, and the epoch is set to 32. We adopt the same data augmentation strategies with the latest methods such as LDF [45], ITSDNet [44] and PFSNet [7]. The prediction results do not need any post-processing.

B. Comparisons with state-of-the-arts

The experimental process involves 16 state-of-the-art methods in the past three years. Four of them in 2019 include BANet [1], BASNet [2], PoolNet [15] and CPD [3]. Seven methods in 2020 include LDF [45], MINet [10], GCPANet [61], GateNet [58], DFI [60], ITSDNet [44] and U2Net [59]. The methods published in 2021 include PurNet [62] PA-KRN [8], VST [14], PSGLoss [51] and PFSNet [7].

Performance and efficiency. The experimental results verify the performance and efficiency of our BBRF. On the one hand, from BBRF in Tab. I and the PR curve in Fig. 7, our method can achieve significant performance improvement on various datasets. For example, the F_β^m of the previous methods on PASCAL-S are mostly at a similar level, the highest is 0.843 of LDF [45], while our method reaches 0.869, a relative increase of 3.1%. Besides, our method has a particularly low MAE (from 0.031 to 0.022 on ECSSD) and a particularly high F_β^* (from 0.876 to 0.891 on PASCAL-S). Based on the observation in Fig. 1, this is most likely caused by the improvement of the segmentation effect of large or small-scale objects. On the other hand, from Tab. II, our BBRF-tiny

TABLE II
COMPARISONS OF PARAMETERS, CALCULATION COST, PEAK MEMORY AND LATENCY BETWEEN OUR PROPOSED BBRF AND THE STATE-OF-THE-ART METHODS.

Method	Crop Size	Params (M)	Macs (G)	Peak Memory (MB)	Latency (MS)
BASNet [2]	256^2	127.4	87.1	2178	17
MINet [10]	320^2	162.4	105.4.	2348	30
GateNet [58]	384^2	128.6	162.1	2286	18
LDF [45]	352^2	25.2	15.5	1318	19
PA-KRN [8]	352^2	102.2	98.7	2110	21
PFSNet [7]	352^2	31.2	45.4	2004	24
Conformer [47]	352^2	81.8	58.4	2384	33
VST [14]	224^2	44.5	23.2	1948	40
BBRF-tiny	224^2	74.4	27.1	1388	30
BBRF	352^2	74.4	46.0	1772	45

TABLE III
ABLATION EXPERIMENTS OF THE BBRF. THE LAST LINE IS THE FINAL RESULT OF OUR METHOD.

ID	Method	ECSSD		DUTS-TE		DUT-OMRON	
		$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$
a	Res-18	.765	.063	.714	.069	.642	.091
b	Swin-T	.846	.032	.845	.033	.775	.052
c	BES	.941	.025	.879	.026	.793	.044
d	BES+DCAM+PPM	.942	.026	.874	.027	.784	.045
e	BES+DCAM+GMS	.944	.024	.881	.026	.798	.044
f	**BES+DCAM+LCS**	**.950**	**.022**	**.893**	**.025**	**.814**	**.042**

with a 224^2 input resolution has mostly surpassed the previous methods on five commonly used metrics. For example, on the DUT-OMRON dataset, the F_β^m metric values of LDF [45] and VST [14] are 0.773 and 0.756, while ours can reach 0.810. It is worth mentioning that an appropriate reduction in the resolution of detail branches may not produce a substantial performance degradation. This is an additional effect of our Bilateral Extreme Stripping framework. The reason for this phenomenon is that the stripped detail branches do not require much down-sampling to obtain semantic information. Therefore, an appropriate reduction in the initial resolution can also extract rich details. This reveals that our method can make a better trade-off between performance and efficiency.

Visual comparison. The visualization results of the proposed BBRF and nine representative state-of-the-art methods are shown in Fig. 6. It is easy to see that the existing methods may have the following problems: 1) Some objects are missing in the prediction result when there are multiple salient objects. 2) Incomplete prediction of a single object. 3) Noise caused by improper handling of details. Among the methods compared, VST is based on the transformer model, while other methods are based on CNN. We observe that VST can handle problems 1 and 2 better, but the segmentation details are often not ideal. On the contrary, the methods based on CNN can better extract details, while they may be difficult to deal with problem 1 and 2 due to the limited receptive field.

As can be seen from the third column, our BBRF has a more comprehensive receptive field and can effectively handle objects with scale variations. 1) For large-scale objects, the proposed method can resolve details better. For example, our salient map in the first row can segment the gap in the aircraft steps. 2) For small-scale including multi-object scenes, our method can find the most significant object from the global view. For instance, the results in the last row show that our method can find the location of the bird on the way and identify the target as a significant object just like the manually labeled results. 3) Our method not only has superior performance in extremely large- and small-scale scenes, but also achieves accurate segmentation for regular-size objects. The results in the middle part of Fig. 6 shows that the separation of details and semantics also generates better segmentation results for regular-size objects.

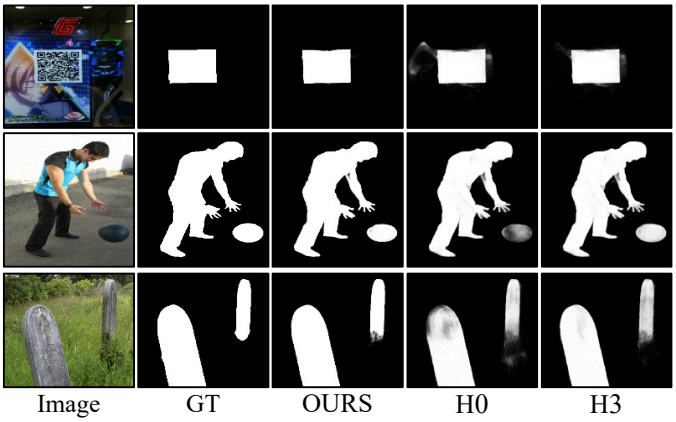

Fig. 8. Visualization of prediction results of different prediction heads in SPD. H0, H3 represent the predicted maps of the first and fourth decision paths, respectively.

Before and after the chain loop correction. In the proposed SPD, error weighting is performed between adjacent decision paths. After chain reinforcement, the features can be improved. Our Loop Compensation Strategy emphasizes feature-level mutual compensation between different decision paths. As shown in the second row of images in Fig. 8, the balloon predicted by the first decision path is ambiguous, while the later paths compensate for more precise results. In fact, we design multiple decision paths to complement each other in the feature dimension, and finally achieve the segmentation of objects with scale variation.

C. Ablation studies

To verify the innovations of this paper, we conduct ablation experiments on the proposed Bilateral Extreme Stripping, Dynamic Complementary Attention Module and Loop Compensation Strategy.

The effect of Bilateral Extreme Stripping encoder. We use different backbones to construct FPN [73] frameworks for performance comparison. As shown in Tab. III, compared to using the CNN (a) or transformer model directly (b), the basic BES encoder (c) can effectively improve the performance. Here, the BES encoder adopts pixel-by-pixel addition to filter features. To make better use of the characteristics of bilateral features, we embed DCAMs (d) into our BES encoder, which can significantly reduce the MAE and improve the F_β^m metric.

TABLE IV
ABLATION STUDY OF TRANSFORMER BRANCH WITH DIFFERENT INPUT RESOLUTIONS.

	Complexity		ECSSD		DUTS-TE		DUT-OMRON	
Crop Size	$Params(M)$	$Mac(G)$	$F_\beta^* \uparrow$	MAE $\downarrow$	$F_\beta^* \uparrow$	MAE $\downarrow$	$F_\beta^* \uparrow$	MAE $\downarrow$
28×28	74.4	35.9	.953	.029	.902	.031	.827	.057
56×56	74.4	46.0	.963	.022	.916	.025	.843	.042
84×84	74.4	63.8	.962	.022	.917	.025	.841	.043

TABLE V
PERFORMANCE AND EFFICIENCY COMPARISON BETWEEN CONFORMER AND OUR BBRF.

	Complexity		ECSSD		DUTS-TE		DUT-OMRON	
Method	$Params(M)$	$Mac(G)$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$
Conformer+LCS	81.8	58.4	.835	.026	.877	.030	.809	.050
BBRF-tiny	**74.4**	**27.1**	.948	.024	.890	.026	.810	**.039**
BBRF	**74.4**	46.0	**.950**	**.022**	**.893**	**.025**	**.814**	.042

BES vs Conformer [47]. We replace BBRF's BES encoder with the Conformer [47] for experimentation. The comparison results with the same configuration and input resolution are shown in Tab. V. It can be seen that our method has advantages in performance and efficiency compared to the Conformer. Specifically, our method reduces the computational cost by 21% (from 58.4 G to 46.0 G), while the performance has been greatly improved. Especially our method can effectively reduce MAE (from 0.030 to 0.025 on DUTS-TE). This result can prove the efficiency and effectiveness of the proposed method. Moreover, our BBRF-tiny is about 2 times faster than Conformer in computation while keeping the performance basically the same.

LCS vs multi-scale modules. We use the same backbone to construct the network in the way shown in Fig. 2 for comparison. The results are listed in Tab. III. Compared with the commonly used multi-scale modules (d, e), LCS (f) has significant performance advantages. Besides, our BES has reduced the MAE to a very low level, while LCS can further improve the effect of the model. This can demonstrate the role of LCS in our BBRF.

The input resolution of semantic branch. Adjusting the input resolution reasonably of the proposed BES encoder can effectively improve the efficiency. As shown in Tab. IV, when the input resolution of the detail branch is fixed, an increase in the resolution of the semantic branch will not always bring a significant performance improvement. Therefore, we choose the settings in the second row to build our model.

D. Exploration on extremely large- / small-scale objects

We propose BBRF to cope with the challenges brought by scale variation, especially extremely large- or small-scale salient objects. To verify the advantages of the proposed method, we set up more detailed experiments.

First, Fig. 1 and Fig. 6 show the advantages of BBRF to handle scale variation. It can be seen that BBRF can achieve the best results on objects with different scales, which reveals the scale robustness of the model.

Second, Tab. VI shows the effect of BBRF in solving asymmetric segmentation requirements. On the one hand, BBRF has a significant performance improvement for small-scale object segmentation. For example, on the DUT-OMRON dataset, the F_β^m increases from 0.629 to 0.679, a relative increase of 7.9%. In contrast, the entire dataset has an average improvement of only 2.6%. On the other hand, BBRF can greatly reduce the MAE of images containing large-scale objects. It is worth noting that the transformer-based method (VST) has a relatively low MAE on large-scale objects. But its performance on small-scale objects is unsatisfactory. CNN-based methods show the opposite trend. Tab. VI verifies that BBRF can take into account the advantages of the both models, which also reflects the flexible receptive fields of our method.

Besides, to understand the contribution of each part more clear, we construct the overall framework BBRF step-by-step. As show in Tab. VII, BES (Swin) builds a higher baseline, which can show excellent segmentation results in extremely large or small-scale object scenes at the same time. This explains the significant effect of BES on extremely large or small-scale object segmentation. Nevertheless, other contributions of BBRF further optimize the performance. For example, when each part is added gradually, the F_β^m of ECSSD-Large gradually reaches 0.968, while the MAE decreases to 0.038. In addition, the experimental results show a similar trend in the five datasets compared. The above experiments could prove the effectiveness of each part of BBRF.

Lastly, to fully verify the performance of our BBRF in extreme-scale scenarios, we built EX-LARGE and EX-SMALL based on existing datasets, which contains more challenging images. Specifically, we merge DUTS-TE [69], HKU-IS [67], ECSSD [66], PASCAL-S [68], DUT-OMRON [65] together, and calculate the pixel ratio of salient objects in each image. And then we sort all the images according to the ratio. We select the first 1,000 large-scale images and their annotations to construct EX-LARGE, and select the last 1,000 small-scale images and their annotations to construct EX-SMALL. As shown in Tab. VIII, our method can significantly reduce the MAE of large-scale objects and the F_β^m of small-scale objects. This highlights that our method can better deal with extreme scale objects. In addition, we will also release these two statistical datasets for further study.

V. LIMITATION

Although the results of the proposed network have reached the current optimal performance, there are still the following points worthy of further exploration. 1) The proposed method is based on the existing datasets, and does not annotate a more explanatory and accurate benchmark consisting of extremely large- or small-scale objects. 2) Our method can better handle scale variation, but the performance on extremely large- or small-scale objects is still relatively low compared to regular-scale objects, which still needs to be further explored. We will continue to explore this issue in our future work.

VI. CONCLUSION

In this paper, we explore that it is difficult to segment large- or small-scale objects due to their asymmetric segmentation

TABLE VI
PERFORMANCE COMPARISON ON LARGE OR SMALL-SCALE OBJECT DATASETS. HERE, LARGE-SCALE OBJECTS ARE DEFINED AS THE TOP 20% OF THE IMAGES IN THE DATASET IN ASCENDING ORDER OF SCALE RATIO. SMALL-SCALE OBJECTS ARE DEFINED AS THE TOP 20% OF THE IMAGES IN THE DATASET SORTED IN DESCENDING ORDER OF THE SCALE RATIO.

	ECSSD				HKU-IS				DUTS-TE				DUT-OMRON				PASCAL-S			
	Large		Small		Large		Small		Large		Small		Large		Small		Large		Small	
Method	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$
MINet [10]	.947	.060	.877	.020	.955	.040	.820	.020	.926	.059	.633	.028	.891	.068	.524	.049	.907	.103	.668	.037
LDF [45]	.946	.063	.893	.020	.953	.040	.842	.020	.925	.058	.715	.022	.879	.072	.590	.040	.906	.101	.724	.032
PFSNet [7]	.941	.055	.899	.017	.956	.037	.843	.019	.925	.055	.691	.028	.893	.069	.573	.047	.882	.108	.715	.035
PA-KRN [8]	.936	.063	.911	**.013**	.954	.039	.859	.019	.919	.061	.756	.020	.883	.071	.629	**.038**	.880	.122	.727	.036
VST [14]	.959	.049	.846	.025	.963	.037	.783	.021	.938	.050	.601	.033	.910	.060	.491	.061	.936	.072	.637	.049
BBRF	**.968**	**.038**	**.926**	.014	**.965**	**.028**	**.897**	**.013**	**.948**	**.039**	**.806**	**.017**	**.922**	**.045**	**.679**	.042	**.933**	**.071**	**.777**	**.030**

TABLE VII
ABLATION STUDY OF OUR LCS, DCAM AND BES ON LARGE OR SMALL-SCALE OBJECT DATASETS. HERE, LARGE-SCALE OBJECTS ARE DEFINED AS THE TOP 20% OF THE IMAGES IN THE DATASET IN ASCENDING ORDER OF SCALE RATIO. SMALL-SCALE OBJECTS ARE DEFINED AS THE TOP 20% OF THE IMAGES IN THE DATASET SORTED IN DESCENDING ORDER OF THE SCALE RATIO.

	ECSSD				HKU-IS				DUTS-TE				DUT-OMRON				PASCAL-S			
	Large		Small		Large		Small		Large		Small		Large		Small		Large		Small	
Method	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$
BES(Swin)	.960	.046	.897	.020	.959	.034	.868	.016	.945	.044	.748	.021	.913	.051	.625	.043	.930	.085	.770	.033
BES(Swin)+DCAM	.963	.044	.901	.017	.961	.033	.870	.013	.946	.043	.751	.019	.917	.050	.628	**.041**	.932	.083	.774	.031
BES(Swin)+DCAM+SPD	.965	.040	.919	.015	.961	.030	.887	.014	.945	.042	.801	.019	.919	.047	.670	.043	.930	.068	**.778**	.031
BES (Swin)+DCAM+SPD+LCS	**.968**	**.038**	**.926**	**.014**	**.965**	**.028**	**.897**	**.013**	**.948**	**.039**	**.806**	**.017**	**.922**	**.045**	**.679**	.042	**.933**	**.071**	.777	**.030**

TABLE VIII
PERFORMANCE COMPARISON ON NEWLY BUILT EX-LARGE AND EX-SMALL DATASETS.

Method	EX-LARGE		EX-SMALL	
	$F_\beta^m \uparrow$	MAE $\downarrow$	$F_\beta^m \uparrow$	MAE $\downarrow$
MINet [10]	.927	.072	.487	.033
LDF [45]	.924	.074	.585	**.028**
PFSNet [7]	.915	.072	.561	.035
PA-KRN [8]	.910	.079	.630	.029
VST [14]	.945	.053	.447	.046
BBRF	**.954**	**.045**	**.684**	.029

requirements. We deconstruct the role of receptive fields in SOD and introduce a Bilateral Extreme Stripping encoder based on the simplified vision transformer and the lightweight CNN for the broader receptive fields. To combine bilateral features and generate a more elastic perceptual perspective, we propose a Dynamic Complementary Attention Module to enhance flexible feature representation. To further highlight the scale-specific perception of different scale ranges, we propose a Loop Compensation Strategy for complementary training of switch-paths through loop chain correction manner. Experiments show that our method can achieve impressive results under scale variable scenes.

VII. ACKNOWLEDGEMENT

This work is partially supported by the National Natural Science Foundation of China under the Grant 62132002 and Grant 62102206.

REFERENCES

[1] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *ICCV*, 2019.

[2] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[3] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3907–3916.

[4] J.-J. Liu, Q. Hou, Z.-A. Liu, and M.-M. Cheng, "Poolnet+: Exploring the potential of pooling for salient object detection," *IEEE TPAMI*, pp. –, 2021.

[5] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3203–3212.

[6] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 678–686.

[7] M. Ma, C. Xia, and J. Li, "Pyramidal feature shrinking for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2311–2318.

[8] B. Xu, H. Liang, R. Liang, and P. Chen, "Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3004–3012.

[9] S. Mohammadi, M. Noori, A. Bahri, S. G. Majelan, and M. Havaei, "Cagnet: Content-aware guidance for salient object detection," *Pattern Recognition*, vol. 103, p. 107303, 2020.

[10] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9413–9422.

[11] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 186–202.

[12] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 660–668.

[13] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 355–370.

[14] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4722–4732.

[15] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *IEEE CVPR*, 2019.

[16] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3085–3094.

[17] Y. Zhao, J. Zhao, J. Li, and X. Chen, "Rgb-d salient object detection with ubiquitous target awareness," *IEEE Transactions on Image Processing*, vol. PP, pp. 1–1, 09 2021.

[18] X. Chai, F. Shao, Q. Jiang, and Y.-S. Ho, "Mstgar: Multioperator-based stereoscopic thumbnail generation with arbitrary resolution," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1208–1219, 2019.

[19] S. Yang, Q. Jiang, W. Lin, and Y. Wang, "Sgdnet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment," in *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2019.

[20] Q. Yao and X. Gong, "Saliency guided self-attention network for weakly-supervised semantic segmentation," *arXiv preprint arXiv:1910.05475*, 2019.

[21] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3089–3098.

[22] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 234–250.

[23] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1741–1750.

[24] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 714–722.

[25] Y. Zeng, M. Feng, H. Lu, G. Yang, and A. Borji, "An unsupervised game-theoretic approach to saliency detection," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4545–4554, 2018.

[26] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4142–4150.

[27] Y. Zeng, H. Lu, L. Zhang, M. Feng, and A. Borji, "Learning to promote saliency detectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1644–1653.

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[29] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 202–211.

[30] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5968–5977.

[31] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7264–7273.

[32] L. Zhu, J. Chen, X. Hu, C.-W. Fu, X. Xu, J. Qin, and P.-A. Heng, "Aggregating attentional dilated features for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3358–3371, 2019.

[33] Y. Xu, D. Xu, X. Hong, W. Ouyang, R. Ji, M. Xu, and G. Zhao, "Structured modeling of joint deep feature and prediction refinement for salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3789–3798.

[34] B. Wang, Q. Chen, M. Zhou, Z. Zhang, X. Jin, and K. Gai, "Progressive feature polishing network for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 128–12 135.

[35] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1623–1632.

[36] Y. Liu, Q. Zhang, D. Zhang, and J. Han, "Employing deep part-object relationships for salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1232–1241.

[37] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[38] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8779–8788.

[39] L. Zhang, J. Zhang, Z. Lin, H. Lu, and Y. He, "Capsal: Leveraging captioning to boost semantics for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6024–6033.

[40] H. Wen, C. Yan, X. Zhou, R. Cong, Y. Sun, B. Zheng, J. Zhang, Y. Bao, and G. Ding, "Dynamic selective network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 9179–9192, 2021.

[41] R. Cong, Y. Zhang, L. Fang, J. Li, Y. Zhao, and S. Kwong, "Rrnet: Relational reasoning network with parallel multiscale attention for salient object detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[42] R. Cong, N. Yang, C. Li, H. Fu, Y. Zhao, Q. Huang, and S. Kwong, "Global-and-local collaborative learning for co-salient object detection," *arXiv preprint arXiv:2204.08917*, 2022.

[43] J. Wei, S. Wang, and Q. Huang, "F^3net: Fusion, feedback and focus for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 321–12 328.

[44] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9141–9150.

[45] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 025–13 034.

[46] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," *arXiv preprint arXiv:2101.11986*, 2021.

[47] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local features coupling global representations for visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 367–376.

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

[49] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[50] S. Yang, G. Lin, Q. Jiang, and W. Lin, "A dilated inception network for visual saliency prediction," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 2163–2176, 2019.

[51] S. Yang, W. Lin, G. Lin, Q. Jiang, and Z. Liu, "Progressive self-guided loss for salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 8426–8438, 2021.

[52] G. Liao, W. Gao, Q. Jiang, R. Wang, and G. Li, "Mmnet: Multi-stage and multi-scale fusion network for rgb-d salient object detection," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2436–2444.

[53] M.-M. Cheng, S.-H. Gao, A. Borji, Y.-Q. Tan, Z. Lin, and M. Wang, "A highly efficient model to study the semantics of salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[54] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[56] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11217. Springer, 2018, pp. 334–349. [Online]. Available: https://doi.org/10.1007/978-3-030-01261-8_20

[57] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3051–3068, 2021.

[58] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 35–51.

[59] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, vol. 106, p. 107404, 2020.

[60] J.-J. Liu, Q. Hou, and M.-M. Cheng, "Dynamic feature integration for simultaneous detection of salient object, edge, and skeleton," *IEEE Transactions on Image Processing*, vol. 29, pp. 8652–8667, 2020.

[61] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 599–10 606.

[62] J. Li, J. Su, C. Xia, and Y. Tian, "Salient object detection with purificatory mechanism and structural similarity loss," *IEEE Transactions on Image Processing*, vol. 30, pp. 6855–6868, 2021.

[63] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.

[64] P. Jaccard, "The distribution of the flora in the alpine zone. 1," *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[65] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3166–3173.

[66] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1155–1162.

[67] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5455–5463.

[68] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280–287.

[69] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 136–145.

[70] Z. Zhao, C. Xia, C. Xie, and J. Li, "Complementary trilateral decoder for fast and accurate salient object detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4967–4975.

[71] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 733–740.

[72] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment Measure for Binary Foreground Map Evaluation," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 698–704.

[73] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

Mingcan Ma is currently pursuing the master's degree with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. His research interests include image parsing and image salient object detection.

Changqun Xia received the Ph.D. degree from the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, in July 2019. He is currently an Assistant Professor with the Peng Cheng Laboratory, China. His research interests include computer vision and image/video understanding.

Chenxi Xie received the Bachelor degree in computer science from the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, in 2021. He is currently a graduate student in Beihang University, Beijing. His research interests include visual attention and computer vision.

Xiaowu Chen (Senior Member, IEEE) received the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2001. His research interests include virtual reality, augmented reality, computer graphics, and computer vision.

Jia Li is currently a Full Professor with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. He received his B.E. degree from Tsinghua University in 2005 and Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences, in 2011. Before he joined Beihang University in 2014, he used to work at Nanyang Technological University, Shanda Innovations, and Peking University. His research is focused on computer vision, multimedia and artificial intelligence, especially the visual computing in extreme environments. He has co-authored more than 110 articles in peer-reviewed top-tier journals and conferences. He also has one Monograph published by Springer and more than 60 patents issued from U.S. and China. He is a Fellow of IET, and senior members of IEEE/ACM/CCF/CIE.