Salient Object Detection With Purificatory Mechanism and Structural Similarity Loss

Jia Li, *Senior Member, IEEE*, Jinming Su, Changqun Xia, Mingcan Ma, and Yonghong Tian, *Senior Member, IEEE*

***Abstract*— Image-based salient object detection has made great progress over the past decades, especially after the revival of deep neural networks. By the aid of attention mechanisms to weight the image features adaptively, recent advanced deep learning-based models encourage the predicted results to approximate the ground-truth masks with as large predictable areas as possible, thus achieving the state-of-the-art performance. However, these methods do not pay enough attention to small areas prone to misprediction. In this way, it is still tough to accurately locate salient objects due to the existence of regions with indistinguishable foreground and background and regions with complex or fine structures. To address these problems, we propose a novel convolutional neural network with purificatory mechanism and structural similarity loss. Specifically, in order to better locate preliminary salient objects, we first introduce the promotion attention, which is based on spatial and channel attention mechanisms to promote attention to salient regions. Subsequently, for the purpose of restoring the indistinguishable regions that can be regarded as error-prone regions of one model, we propose the rectification attention, which is learned from the areas of wrong prediction and guide the network to focus on error-prone regions thus rectifying errors. Through these two attentions, we use the *Purificatory Mechanism* to impose strict weights with different regions of the whole salient objects and purify results from hard-to-distinguish regions, thus accurately predicting the locations and details of salient objects. In addition to paying different attention to these hard-to-distinguish regions, we also consider the structural constraints on complex regions and propose the *Structural Similarity Loss*. The proposed loss models the region-level pair-wise relationship between regions to assist these regions to calibrate their own saliency values. In experiments, the proposed purificatory mechanism and structural similarity loss can both effectively improve the performance, and the proposed approach outperforms 19 state-of-the-art methods on six datasets with a notable margin. Also, the proposed method is efficient and runs at over 27FPS on a single NVIDIA 1080Ti GPU.**

***Index Terms*— Salient object detection, purificatory mechanism, error-prone region, structural similarity.**

I. Introduction

VISUAL saliency plays an essential role in the human vision system, which guides human beings to look at the most important information from visual scenes and can be well referred to as the allocation of cognitive resources on information [1], [2]. To model this mechanism of visual saliency, there are two main research branches in computer vision: fixation prediction [3] and salient object detection [4]. This work focuses on the second one (*i.e.*, salient object detection, abbreviated as SOD), which aims to detect and segment the most visually distinctive objects. Over the past years, SOD has made significant progress, and it is also used as an important preliminary step for various vision tasks, such as object recognition [5], tracking [6] and image parsing [7].

To address the SOD task, lots of learning-based methods [10]–[22] have been proposed in recent years, achieving impressive performance on existing benchmark datasets [8], [23]–[27]. However, there still exist two difficulties that hinder the development of SOD. First, it is hard to distinguish these regions with similar foreground and background. As shown in Fig. 1(a)(b), these regions usually confuse the models to cause wrong predictions, and we named these regions as "error-prone region" of models. Second, it is difficult to restore the complex or fine structures. As displayed in Fig. 1(c)(d), the complex structures (*e.g.*, caused by complex illumination and color) and fine structures (*e.g.*, hollows) make it difficult to maintain the structural integrity and clarity. These two problems are especially difficult to deal with for existing SOD methods and greatly hinder the performance of SOD. Due to these difficulties, SOD remains a challenging vision task.

To deal with the first difficulty, some methods [28]–[33] adopt attention mechanisms to weight the features adaptively to focus on salient regions. For examples, Zhang *et al.* [34] introduced an attention guided network to integrate multi-level contextual information by utilizing global and local attentions, consistently improving saliency detection performance. Chen *et al.* [29] proposed the reverse attention to guide the side-output residual learning in a top-down manner to restore the salient object parts and details. For these methods,

Manuscript received December 9, 2019; revised November 5, 2020, February 12, 2021, and March 10, 2021; accepted July 3, 2021. Date of publication July 28, 2021; date of current version August 3, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61922006 and Grant 62088102. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaoming Liu. *(Corresponding authors: Changqun Xia; Jia Li.)*

Jia Li and Mingcan Ma 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 the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: jiali@buaa.edu.cn).

Jinming Su is 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 the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: xiachq@pcl.ac.cn).

Yonghong Tian is with the Department of Computer Science and Technology, Peking University, Beijing 100871, China, and also with the Peng Cheng Laboratory, Shenzhen 518000, China.

Digital Object Identifier 10.1109/TIP.2021.3099405

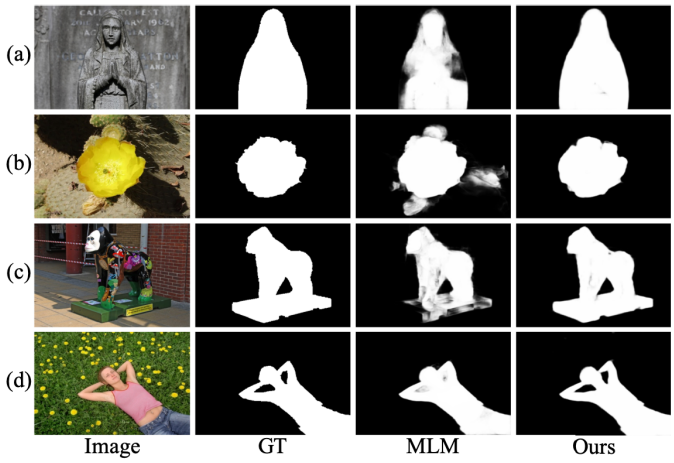

Fig. 1. Difficulties that hinder the development of SOD. In (a)(b), these usually exist regions with similar foreground and background, which confuses the models to cause wrong predictions. In (c)(d), complex structures caused by complex illumination or color and fine hollows make it difficult to maintain the structural integrity and clarity. Images and ground-truth masks (GT) are from ECSSD [8]. Results are generated by MLM [9] and our approach.

although different forms of features are effectively aggregated, the overall goal is to make the prediction results approach ground-truth masks with as larger an intersection as possible, which improves the accuracy of the area that is easy to predict. However, these methods mainly focuses on improving the correctness of large predictable areas, but don't pay enough attention to small error-prone areas. To address the second problem, methods [9], [22], [33], [35]–[39] consider to solve the problem of inaccurate boundaries. For example, Wang *et al.* [35] proposed a local boundary refinement network to recover object boundaries by learning the local contextual information for each spatial position. Wu *et al.* [9] also adopted the foreground contour and edge to guide each other, thereby leading to precise foreground contour prediction and reducing the local noises. In these methods, some special boundary branches and losses are proposed to attend boundaries or local details. In this way, these methods mainly take account of the unary supervision to deal with the complex and fine structures. But for many complex nad fine structures that are influenced by the context, it is difficult to accurately restore only considering the unary information, which only considers the correlation at the pixel level but not at the regional level.

Inspired by these observations and analyses, we propose a novel convolutional neural network with purificatory mechanism and structural similarity loss for image-based SOD. In the network, we propose the purificatory mechanism to purify salient objects by promoting predictable regions and rectifying indistinguishable regions. In this mechanism, we first introduce a simple but effective promotion attention based on spatial and channel attention mechanisms to provide the promotion ability, which assists to locate preliminary salient objects. Next, we propose a novel rectification attention, which predicts the error-prone areas and guides the network to pay more attention to these areas to rectify errors from the aspect of features and losses. These two attentions are used to impose strict weights with different regions of the whole salient objects and formed the purificatory mechanism. In addition, in order to better restore the complex or fine structures of salient objects, we propose a novel structural similarity loss to model and constrain the structural relation on complex regions for better calibrating the saliency values of regions, which can be regarded as an effective supplement to the pixel-level unary constraint. The purificatory mechanism and structural similarity loss are integrated in a progressive manner to pop-out salient objects. Experimental results on six public benchmark datasets verify the effectiveness of our method which consistently outperforms 19 state-of-the-art SOD models with a notable margin. Moreover, the proposed method is efficient and runs at about 27FPS on a single NVIDIA 1080Ti GPU.

The main contributions of this paper include:

1) we propose a novel *Purificatory Mechanism*, which purifies salient objects by promoting predictable regions and rectifying indistinguishable regions;
2) we introduce a novel *Structural Similarity Loss* to restore the complex or fine structures of salient objects, which constrains region-level pair-wise relationship between regions to be as a supplement to the pixel-level unary constraints, assisting regions to calibrate their own saliency values;
3) we conduct comprehensive experiments and the results verify the effectiveness of our proposed method which consistently outperforms 19 state-of-the-art algorithms on six datasets with a fast prediction.

The rest of this paper is organized as follows: Section II reviews the recent development of salient object detection, attention-based SOD methods and boundary-aware SOD methods. Section III presents the purificatory network in detail. Section IV presents the proposed structural similarity loss. In Section V, we evaluate the proposed model, and compare it with the state-of-the-art methods to validate the effectiveness of the model. We conclude the paper in Section VI.

II. Related Work

In this section, we review the related works in three aspects. At the beginning, some representative salient object detection methods are introduced. Next, we present attention mechanisms and attention-based SOD methods. Next, we review the boundary-aware SOD methods.

A. Salient Object Detection

Hundreds of image-based SOD methods have been proposed in the past decades. Early methods mainly adopted hand-crafted local and global visual features as well as heuristic saliency priors such as color difference [40], distance transformation [41] and local/global contrast [42], [43]. More details about the traditional methods can be found in the survey [4].

With the development of deep learning, many deep neural networks (DNNs) based methods [10]–[22] have been proposed for SOD. Lots of deep models are devoted to fully utilizing the feature integration to enhance the performance of DNNs. For example, Lee *et al.* [10] proposed to compare

the low-level features with other parts of an image to form a low-level distance map. Then they concatenated the encoded low-level distance map and high-level features extracted by VGG [44], and connect them to a DNN-based classifier to evaluation the saliency of a query region. Liu *et al.* [11] presented a DHSNet that first made a coarse global prediction by learning various global structured saliency cures and then adopted a recurrent convolutional neural network to refine the details of saliency maps by integrating local contexts step by step, which worked in a global to local and coarse to fine manner.

In addition, Hou *et al.* [12] introduced short connections to the skip-layer structures, which provided rich multi-scale feature maps at each layer, performing salient object detection. Luo *et al.* [13] proposed a convolutional neural network by combining global and local information through a multi-resolution 4×5 grid structure to simplify the model architecture and speed up the computation. Zhang *et al.* [14] adopted a framework to aggregate multi-level convolutional features into multiple resolutions, which were then combined to predict saliency maps in a recursive manner. Wang *et al.* [15] proposed a pyramid pooling module and a multi-stage refinement mechanism to gather contextual information and stage-wise results, respectively. Zhang *et al.* [16] utilized the deep uncertain convolutional features and proposed a reformulated dropout after specific convolutional layers to construct an uncertain ensemble of internal feature units. Chen *et al.* [17] incorporated human fixation with semantic information to simulate the human annotation process to form two-stream fixation-semantic CNNs, which were fused by an inception-segmentation module. Zhang *et al.* [18] proposed a novel bi-directional message passing model to integrate multi-level features for SOD.

These methods usually integrate multi-scale and multi-level feature by complex structures to improve the representation ability of DNNs. To simply and effectively integrate these features, we add lateral connections to transfer encoded features to assist the decoder and adopt a top-down architecture to propagate high-level semantics to low-level details as guide of locating salient objects as well as restoring object details.

B. Attention-Based Methods

Attention mechanism of DNNs is inspired from human perception process, which weights the features to encourage one model to focus on important information. The mechanism was first applied in machine translation [45] and then widely used in the field of computer vision due to its effectiveness. For example, Mnih *et al.* [46] applied an attention-based model to image classification tasks. In [47], SCA-CNN that incorporated spatial and channel-wise attention mechanisms in a CNN are proposed to modulate the sentence generation context in multi-layer feature maps, encoding where and what the visual attention is, for the task of image captioning. Chu *et al.* [48] combined the holistic attention model focusing on the global consistency and the body part attention model focusing on detailed descriptions for human pose estimation. Fu *et al.* [49] proposed the dual attention network that adopted the position attention module aggregated the feat at each position and the channel attention module emphasizes interdependent channel maps for scene segmentation. Woo *et al.* [50] proposed Convolutional Block Attention Module (CBAM) to efficiently help the information flow within the network by learning which information to modality or suppress.

Due to the effectiveness of attention mechanisms for feature enhancement, they are also applied to saliency detection. Liu *et al.* [28] proposed a pixel-wise contextual attention network to learn to attend to informative context locations for each pixel by two attentions: global attention and local attention, guiding the network learning to attend global and local contexts, respectively. Feng *et al.* [38] designed the attentive feedback modules to control the message passing between encoder and decoder blocks, which was considered an opportunity for error corrections. Zhang *et al.* [30] leveraged captioning to boost semantics for salient object detection and introduced a textual attention mechanism to weight the importance of each word in the caption. In [31], a holistic attention module was proposed to enlarge the coverage area of these initial saliency maps since some objects in complex scenes were hard to be completely segmented. Zhao and Wu [32] presented a pyramid feature attention network to enhance the high-level context features and the low-level spatial structural features. Wang *et al.* [33] proposed a pyramid attention structure to offer the representation ability of the corresponding network layer with an enlarged receptive field.

In the above methods, attention mechanisms (spatial attention and channel attention) are used to enhance the localization and awareness of salient objects. These attentions play good roles in promoting feature attention to salient regions, but lacks attention to small regions prone to mis-prediction. Unlike these methods, we propose the purificatory mechanism, which introduce two novel attentions: promotion attention and rectification attention. The first attention is dedicated to promoting the feature representation of salient regions, while the second one is dedicated to rectifying the features of error-prone regions.

C. Boundary-Aware Methods

Some methods [9], [22], [33], [35]–[39] consider that unclear object boundaries and inconsistent local details are important factors affecting performance of SOD. Li *et al.* [36] considered contours as useful priors and proposed to facilitate feature learning in SOD by transferring knowledge from an existing contour detection model. In [37], an edge detection branch was used to assist the deep neural network to further sharpen the details of salient objects by joint training. Feng *et al.* [38] presented a boundary-enhanced loss for learning fine boundaries and worked with the cross-entropy loss for saliency detection. Qin *et al.* [39] also proposed a loss for boundary-aware SOD and the loss guided the network to learn in three levels: pixel-level, patch-level and map-level. Besides, more effective loss functions, such as meas intersection-over–union loss. weighted binary cross-entropy loss and affinity field matching loss, have been made to capture the quality factors for salient object detection tasks [51]. In [33], a salient edge detection module is introduced to emphasize on the importance of salient edge information, encouraging

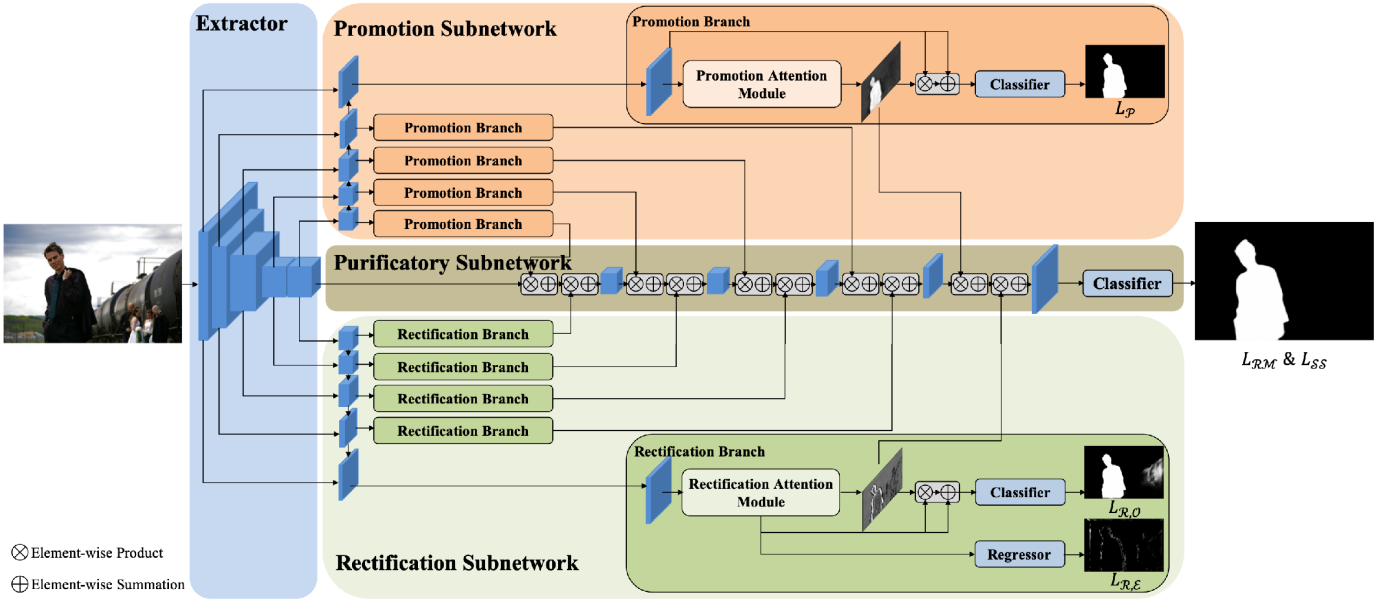

Fig. 2. The framework of our approach. We first extract the common features by extractor, which provides the features for the other three subnetworks. In detail, the promotion subnetwork produces promotion attention to guide the model to focus on salient regions, and the rectification subnetwork give the rectification attention for rectifying the errors. These two kind of attentions are combined to formed the purificatory mechanism, which is integrated in the purificatory subnetwork to refine the prediction of salient objects progressively.

better edge-preserving SOD. And Su *et al.* [22] proposed a boundary-aware network, which split salient objects into boundaries and interiors, extracted features from different regions to ensure the representation of each region, and then fused to obtain good results.

These methods usually utilize some special boundary branch and loss to attend boundaries or local details. But for many complex and fine structures that are influenced by the context, it is difficult to accurately restore only considering the unary information. Our method differs with these methods by introducing the structural similarity loss, which models and constrains the pair-wise structural relation on complex regions for better calibrating the saliency values of regions and is an effective supplement to the pixel-level unary constraint.

III. Purificatory Network

To address these problems (*i.e.*, indistinguishable regions and complex structures), we propose a novel purificatory network (denoted as **PurNet**) for SOD. In this method, different regions are attended by corresponding attentions, *i.e.*, promotion attention and rectification attention. The first one is to promote attention in salient regions and the second one aims to rectify errors for salient regions. In terms of the architecture, the network includes four parts: the feature extractor, the promotion subnetwork, the rectification subnetwork and the purificatory subnetwork. In this section, we first overview the whole purificatory network and then introduce each part separately. Details of the proposed approach are described as follows.

A. Overview

A diagram of the top-down architecture with feature transferring and utilization is as shown in Fig. 3, the proposed PurNet has a top-down basic architecture with lateral connections, which is used by the feature pyramid network (FPN) [52] based on the encoder-decoder form. In our method, PurNet consists of four parts, and the first part (*i.e.*, the extractor) provides the common features for the other three ones (regarded as decoders). Each of the rest three parts forms an encoder-decoder relation with the feature extractor, and decodes the received features respectively. In the rest three decoders, the promotion subnetwork is used to provide the promotion features which is utilized to improve the localization ability and semantic information for salient regions and the rectification subnetwork provides rectification features which can provide the rectification attention for rectifying the errors, while the purificatory subnetwork uses the purificatory mechanism to refine the prediction of SOD progressively.

B. Feature Extractor

To see the Fig. 3, the purificatory network tasks ResNet-50 [53] as the feature extractor, which is modified to remove the last global pooling and fully connected layers for the pixel-level prediction task. Feature extractor has five residual modules for encoding, named as $\mathcal{E}_1(\pi_1), \ldots, \mathcal{E}_5(\pi_5)$ with parameters $\pi_1, \ldots, \pi_5$. To obtain larger feature maps, the strides of all convolutional layers belonging to last residual modules $\mathcal{E}_5$ are set to 1. To further enlarge the receptive fields of high-level features, we set the dilation rates [54] to 2 and 4 for convolution layers in $\mathcal{E}_4$ and $\mathcal{E}_5$, respectively. For a $H \times W$ input images, a $\frac{H}{16} \times \frac{W}{16}$ feature map is output by the feature extractor.

In order to integrate multi-level and multi-scale features, we adopt lateral connections to transfer the features of each encoding module to the decoder by a convolution layer with 128 kernels of 1×1, which also compresses the channels of high-level features for later processing and integration.

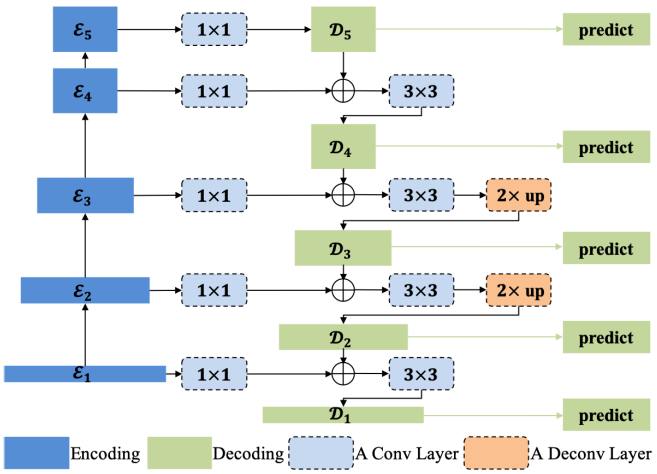

Fig. 3. The backbone of feature extractor. We adopt five residual modules for encoding, lateral connections to transfer the features of each encoding module to the decoder by a convolution layer with 128 kernels of 1×1 for utilizing multi-level and multi-scale features, and convolution layers with 128 kernels of 3×3 followed by a $2\times$ upsampling deconvolution layer for decoding and restoring features.

In addition, we use a top-down architecture to propagate high-level semantics to low-level details as guide of locating salient objects as well as restoring object details. In this architecture, features from same-level encoding feature and higher-level decoding features are added, and a convolution layer with 128 kernels of 3×3 is used to decode these features. We use learnable deconvolution to perform $2\times$ upsampling to align and restore features.

For the following three subnetworks (*i.e.*, the promotion, rectification and purificatory subnetworks), there is a set of learned decoding features $\mathcal{D}_i, i \in \{1, \ldots, 5\}$, respectively. The three subnetworks mainly process these decoding features and predict the corresponding expected results.

C. Promotion Subnetwork

1) Promotion Attention: In general, when there are some distractions in the background, the location of salient objects is difficult to be detected as shown in Fig. 1(a)(b). Some methods [28], [29], [34] consider to make their models focus on the salient regions by spatial attention and channel attention mechanisms. In these two mechanisms, the former can be used to enhance the localization capability, and the latter aims to enhance semantic information [32]. For example,CBAM [50] and PAGRN [34] adopt the cascade way to reconcile spatial and channel information and have has proven to be effective. However, this way emphasizes the sequence of spatial and channel information in transmission, which will cause the loss of information in some complex scenes. In order to capture the contextual information in spatial and channel dimension, we pay more attention to the balance and reinforce of independent spatial information and channel information. Therefore, we propose a simple but effective parallel structure to provide the promotion ability.

We present the structure of the promotion attention module as depicted in Fig. 4. This moduel is based on existing

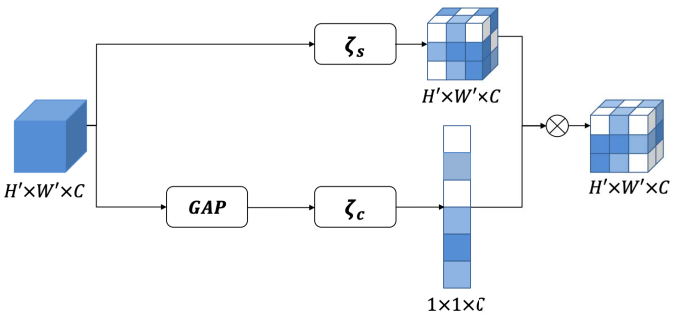

Fig. 4. The structure of the promotion attention module. The Softmax operation on spatial dimension (ζ_s) is used to extract the spatial attention, and global average pooling (GAP) followed by the channel Softmax operation (ζ_c) is used to obtain channel attention. The two attentions are multiplied as the promotion attention.

spatial and channel attention without additional parameters. We denote input convolutional features as $\mathrm{F}_{\mathcal{P}} \in \mathbb{R}^{H'\times W'\times C}$. The promotion attention is generated as follows:

$$\mathrm{A}_{\mathcal{P}} = \zeta_s(\mathrm{F}_{\mathcal{P}}) \otimes \zeta_c(GAP(\mathrm{F}_{\mathcal{P}})), \tag{1}$$

where $\zeta_s(\cdot)$ and $\zeta_c(\cdot)$ denotes the Softmax operation on the spatial and channel dimension respectively, $GAP(\cdot)$ is the operation of global average pooling, and $\otimes$ represents element-wise product.

In Eq. (1), the first item $\zeta_s(\mathrm{F}_{\mathcal{P}})$ is spatial attention, where a Softmax operation on spatial dimension is directly conducted to obtain the spatial weights, and the second item $\zeta_c(GAP(\mathrm{F}_{\mathcal{P}}))$ is the channel attention, where global average pooling is adopted to remove the effect of spatial for getting a vector of length C followed by a Softmax operation on channel dimension to obtain the channel weights. Then, the attentions of spatial and channel dimension decouple and they are integrated by an element-wise product operation. In this manner, the advantage of our parallel structure lies in the adaptive allocation of spatial and channel information weights, thus avoiding artificial design and interference of different information weights and leading to locate preliminary salient objects more efficiently. Some visual examples can be found in the third column of Fig. 5.

2) Subnetwork: As shown in Fig. 2, the promotion attention module exists in the promotion subnetwork. In the promotion subnetwork, features from the five lateral connections of the feature extractor are firstly decoded. And then, each branch processes one of different level decoding features. For each branch, we represent input decoding convolutional features as $\mathrm{F}_{\mathcal{P}} \in \mathbb{R}^{H\times W\times C}$ (the same features $\mathrm{F}_{\mathcal{P}}$ in Eq. (1)). Then the promotion attention module is utilized to weight the input features $\mathrm{F}_{\mathcal{P}}$ by the following operation:

$$\mathrm{M}_{\mathcal{P}} = \mathrm{F}_{\mathcal{P}} \otimes \mathrm{A}_{\mathcal{P}} + \mathrm{F}_{\mathcal{P}}. \tag{2}$$

The generated features $\mathrm{M}_{\mathcal{P}}$ is then classified by a classifier, which includes two convolution layers with 128 kernels of 3×3 and 1×1, and one kernel of 1×1 followed by a Sigmoid and upsampling operation.

For the sake of simplification, these five branches of the promotion subnetwork are denoted as $\phi^{(i)}_{\mathcal{P}}(\pi^{(i)}_{\mathcal{P}}) \in [0,1]^{H\times W\times 1}, i \in \{1, \ldots, 5\}$, where $\pi^{(i)}_{\mathcal{P}}$ is the set of

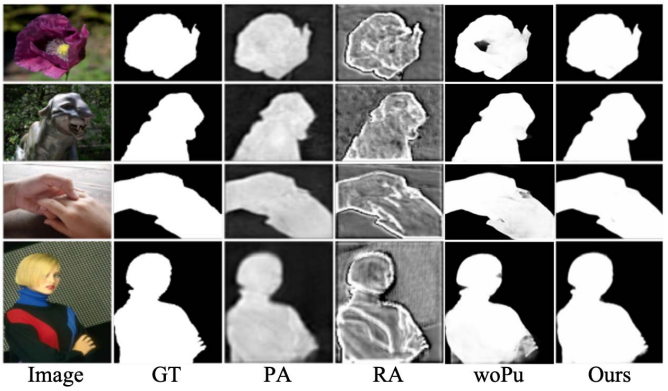

Fig. 5. Visual examples of the purificatory mechanism. GT: ground-truth mask, PA: the promotion attention, RA: the rectification attention, woPu: the prediction without purificatory mechanism, Ours: prediction of our approach.

parameters of $\phi_{\mathcal{P}}^{(i)}$. As mentioned earlier, the promotion subnetwork aims to learn the promotion attention. To achieve this, we expect the output of the promotion attention module to approximate the ground-truth masks of SOD (represented as G) by minimizing the loss:

$$L_{\mathcal{P}} = \sum_{i=1}^{5} BCE(\phi_{\mathcal{P}}^{(i)}(\pi_{\mathcal{P}}^{(i)}), G), \tag{3}$$

where $BCE(\cdot,\cdot)$ means the binary cross-entropy loss function with the following formulation:

$$BCE(P, G) = -\sum_{i}^{H\times W} (G_i \log P_i + (1 - G_i)\log(1 - P_i)), \tag{4}$$

where P_i and G_i represents the ith pixel of predicted maps and ground-truth masks of salient objects, respectively.

By taking multi-level lateral features from feature extractor as input, the promotion subnetwork can learning the promotion attention in multi-scale manner, which is fed to the purificatory subnetwork to promote attention to salient regions and demonstrates the strong promotion ability for SOD.

D. Rectification Subnetwork

In order to restore the structure of confusing or complicated areas (these areas can be regarded as error-prone regions of one model), we present the rectification mechanism, which is obtained by predicting the error-prone regions of the model. We are paying more attention to these areas and at the same time imposing stricter constraints, thus rectifying these errors.

1) Rectification Attention: As shown in Fig. 1(a)(b), it is difficult to accurately define the attributes and locations of some error-prone areas (*e.g.*, salient regions confused with background). Therefore, we propose the rectification attention to guide the model to focus on these error-prone areas for error correction.

The structure of rectification attention module is shown in Fig. 6. This module exists in the rectification branch. We represent the input features as $\mathrm{F}_{\mathcal{R}} \in \mathbb{R}^{H\times W\times C}$. Then two parallel convolution branches are used to process

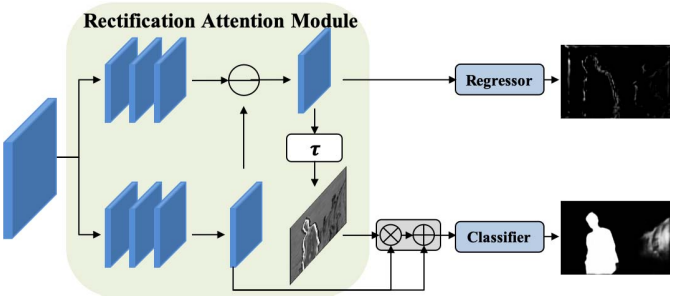

Fig. 6. The structure of the rectification branch. $\tau(\cdot)$ is the Tanh function. The output of the Classifier predicts salient objects and the one of the Regressor predicts errors in the saliency prediction.

the input features, where each branch has two convolution layers with 128 kernels of 3×3 followed by a convolution layer with one kernel of 1×1. We denote the outputs of these two branches as $\mathrm{F}_{\mathcal{R},\mathcal{G}}$ and $\mathrm{F}_{\mathcal{R},\mathcal{O}}$, which mean features of gross regions and object regions (named as gross branches and object branches). The gross features represent potential comprehensive features, while object feature represents predictable features in the object body, and their difference represents mispredicted features. Therefore, we use the subtraction ($\mathrm{F}_{\mathcal{R},\mathcal{E}}$) of $\mathrm{F}_{\mathcal{R},\mathcal{G}}$ and $\mathrm{F}_{\mathcal{R},\mathcal{O}}$ to be as the features of error-prone regions. Next, the rectification attention is generated as follows:

$$\mathrm{A}_{\mathcal{R}} = \tau(\mathrm{F}_{\mathcal{R},\mathcal{E}}), \tag{5}$$

where $\mathrm{F}_{\mathcal{R},\mathcal{E}} = \mathrm{F}_{\mathcal{R},\mathcal{G}} - \mathrm{F}_{\mathcal{R},\mathcal{O}}$ and $\tau(\cdot)$ is the Tanh function, which maps the features into range of $[-1, 1]$ to obtain the rectification attention. The rectification provides the attention to error-prone regions, which are important but almost undiscovered information for SOD. Some examples of rectification attention are shown in the forth column of Fig. 5.

2) Subnetwork: Similar to the promotion attention module, the rectification attention module exists in the rectification subnetwork as shown in Fig. 6. In the subnetwork, features from the five lateral connections of the feature extractor are decoded and as the input to each rectification branch in a multi-level manner. For each branch, the rectification attention is used to weight the object features $\mathrm{F}_{\mathcal{R},\mathcal{O}}$ as follows:

$$\mathrm{M}_{\mathcal{R},\mathcal{O}} = \mathrm{F}_{\mathcal{R},\mathcal{O}} \otimes \mathrm{A}_{\mathcal{R}} + \mathrm{F}_{\mathcal{R},\mathcal{O}}. \tag{6}$$

The generated features $\mathrm{M}_{\mathcal{R},\mathcal{O}}$ is fed to a classifier, which is the same as the classifier in the promotion subnetwork. We denote the object outputs of rectification subnetwork as $\phi_{\mathcal{R},\mathcal{O}}^{(i)}(\pi_{\mathcal{R},\mathcal{O}}^{(i)}) \in [0, 1]^{H\times W\times 1}, i \in \{1, \ldots, 5\}$, where $\pi_{\mathcal{R},\mathcal{O}}^{(i)}$ is the set of parameters of $\phi_{\mathcal{R},\mathcal{O}}^{(i)}$, consisting of the parameters of decoding convolution layer and object branches. And the outputs of classifiers are expected to approximate the ground-truth masks of SOD. The minimizing optimization objective is as follows:

$$L_{\mathcal{R},\mathcal{O}} = \sum_{i=1}^{5} BCE(\phi_{\mathcal{R},\mathcal{O}}^{(i)}(\pi_{\mathcal{R},\mathcal{O}}^{(i)}), G). \tag{7}$$

In addition, an additional regressor is added to the error-prone features $\mathrm{F}_{\mathcal{R},\mathcal{E}}$. The regressor consists of two

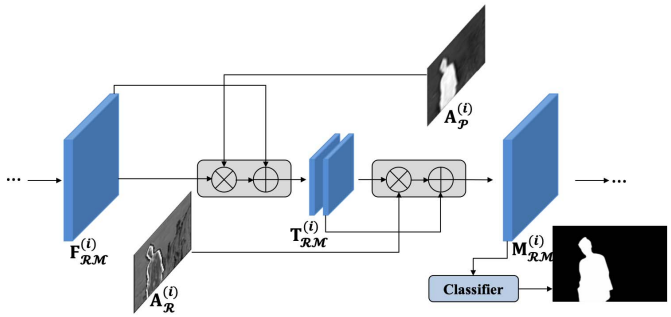

Fig. 7. The part structure of the purificatory subnetwork. The purificatory subnetwork integrates the promotion and rectification attentions by the purificatory mechanism.

convolution layers with 128 kernels of 3×3 and 1×1, and one kernel of 1×1 followed by a Tanh operation. The outputs are the error-prone prediction of rectification subnetwork, denoted as $\phi^{(i)}_{\mathcal{R},\mathcal{E}}(\pi^{(i)}_{\mathcal{R},\mathcal{E}}) \in [-1,1]^{H\times W\times 1}, i \in \{1,\ldots,5\}$, where $\pi^{(i)}_{\mathcal{R},\mathcal{E}}$ is the set of parameters of $\phi^{(i)}_{\mathcal{R},\mathcal{E}}$. The outputs of regressors aim to approximate the error maps of $\phi^{(i)}_{\mathcal{R},\mathcal{O}}(\pi^{(i)}_{\mathcal{R},\mathcal{O}})$ and the error map is defined as $G^{(i)}_{\mathcal{E}} = G - \phi^{(i)}_{\mathcal{R},\mathcal{O}}(\pi^{(i)}_{\mathcal{R},\mathcal{O}})$. Obviously, the value of $P^{(i)}_{\mathcal{E}}$ is in the range of $[-1,1]$. In order to learning the error map, we drive predicted error map $P^{(i)}_{\mathcal{E}} = \phi^{(i)}_{\mathcal{R},\mathcal{E}}(\pi^{(i)}_{\mathcal{R},\mathcal{E}})$ to approach its ground truth $G^{(i)}_{\mathcal{E}}$ by minimizing the KL-divergence:

$$L_{\mathcal{R},\mathcal{E}} = \sum_{i=1}^{5} KL(N(G^{(i)}_{\mathcal{E}})||N(P^{(i)}_{\mathcal{E}})), \tag{8}$$

where $KL(\cdot||\cdot)$ means the KL-divergence with the following formulation:

$$KL(G||P) = \sum_{i}^{H\times W} G_i \log\frac{G_i}{P_i}, \tag{9}$$

where P_i and G_i are the ith pixel of the predicted error map $P^{(i)}_{\mathcal{E}}$ and the ground-truth error map of $G^{(i)}_{\mathcal{E}}$, respectively. Also, $N(\cdot)$ in the above equation is a normalization operation, which casts the $G_{\mathcal{E}}$ and $P_{\mathcal{E}}$ into the range [0, 1]. In our method, we add 1 to the input and divide by 2 as the $N(\cdot)$ operation.

Through these operations, the rectification subnetwork provides the rectification attention and predicted error maps to the purificatory subnetwork, which drives PurNet to focus on the error-prone regions and rectify the wrong prediction.

E. Purificatory Subnetwork

1) Usage of Promotion and Rectification Attention: Similar to the promotion subnetwork and rectification subnetwork, the purificatory subnetwork processes the features from feature extractor in a top-down manner, which can refine the SOD prediction progressively.

In our approach, the body of salient objects are first promoted with the help of the promotion attention, and then the error-prone regions of salient objects are rectified by the aid of then the rectification attention. Therefore, these two attentions are combined to purify the salient objects. The purificatory mechanism is integrated in the purificatory subnetwork, the structure of which is shown in 7. For ith decoding stage, the input features $\mathrm{F}^{(i)}_{\mathcal{RM}} \in \mathbb{R}^{H\times W\times C}$ are firstly weighted by the promotion attention by the operation:

$$\mathrm{T}^{(i)}_{\mathcal{RM}} = \mathrm{F}^{(i)}_{\mathcal{RM}} \otimes \mathrm{A}^{(i)}_{\mathcal{P}} + \mathrm{F}^{(i)}_{\mathcal{RM}}. \tag{10}$$

And then a convolution layer with 128 kernels of 3×3 are used to convolve the features to be $\mathrm{T}'^{(i)}_{\mathcal{RM}}$. Next, the rectification attention is fed to weight the produced features as follows:

$$\mathrm{M}^{(i)}_{\mathcal{RM}} = \mathrm{T}'^{(i)}_{\mathcal{RM}} \otimes \mathrm{A}^{(i)}_{\mathcal{R}} + \mathrm{T}'^{(i)}_{\mathcal{RM}}. \tag{11}$$

The generated features $\mathrm{M}^{(i)}_{\mathcal{RM}}$ is input to a classifier, which is the same as the classifier in the promotion subnetwork with two convolution layers with 128 kernels of 3×3 and 1×1, and one kernel of 1×1 followed by a Sigmoid and upsampling operation.

We represent the outputs of the purificatory subnetwork as $\phi^{(i)}_{\mathcal{RM}}(\pi^{(i)}_{\mathcal{RM}}) \in [0,1]^{H\times W\times 1}, i \in \{1,\ldots,5\}$, where $\pi^{(i)}_{\mathcal{RM}}$ is the set of parameters of $\phi^{(i)}_{\mathcal{RM}}$, consisting of the parameters of decoding convolution layer and layers of each stage. And the outputs of classifiers are expected to approximate the ground truths of SOD. The loss is formed by the following operation:

$$L_{\mathcal{RM}} = \sum_{i=1}^{5} IBCE(\phi^{(i)}_{\mathcal{RM}}(\pi^{(i)}_{\mathcal{RM}}), G, P^{(i)}_{\mathcal{E}}), \tag{12}$$

where $P^{(i)}_{\mathcal{E}} = \phi^{(i)}_{\mathcal{R},\mathcal{E}}(\pi^{(i)}_{\mathcal{R},\mathcal{E}})$ represents the error maps and $IBCE(\cdot,\cdot,\cdots)$ means the improved binary cross-entropy loss function with error map from the rectification subnetwork. We give the definition in Section III-E.2. To provide more comprehensive visualization to prove the effectiveness of the proposed purificatory mechanism, we adopt the element-wise sum operation to combine these two features. Some examples without purificatory mechanism are shown in the fifth column of Fig. 5.

2) Improved Loss Function: The predicted can be used to penalize the error-prone areas of the predicted saliency map in the purificatory subnetwork. By the extra constraints, the error-prone areas in the final prediction can be better refined. Toward this end, we propose to optimize the saliency maps to approximate the ground-truth masks of SOD by minimizing the improved binary cross-entropy loss (see Eq. (12)). And the improved loss is defined as follows:

$$IBCE(P,G,E) = -\sum_{i}^{H\times W} (G_i \log P_i + (1-G_i)\log(1-P_i)) \cdot(1+|E_i|), \tag{13}$$

where E_i represents the ith pixel of predicted error maps E and $|\cdot|$ indicates the absolute value operation. In our improved loss, the cross-entropy loss function at each pixel is weighted by the predicted error map, which penalizes the error-prone areas with bigger loss to rectify possible errors.

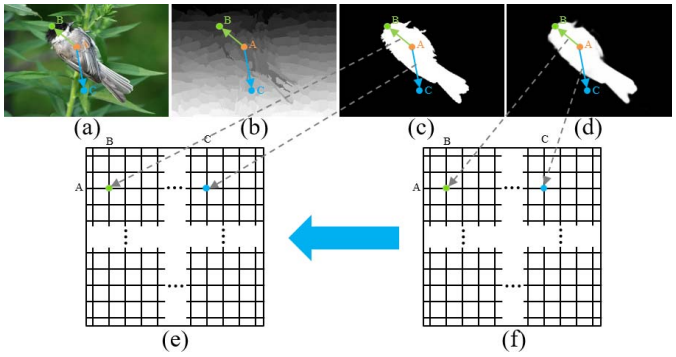

Fig. 8. Construction of the structural similarity matrix. (a) image, (b) super-pixel, (c)ground truth, (d) saliency map of our approach, (e)(f)structural matrices of the ground truth and saliency map.

IV. Structural Similarity Loss

Through the purificatory mechanism, different regions (*i.e.*, simple regions and error-prone regions) of salient objects are processed and the performance is greatly improved. In addition to paying different attention to these indistinguishable regions, we also consider the structural constraints on complex regions as useful information for salient object detection. Toward this end, we propose a novel structural similarity loss (as shown in Fig. 8) to constrain the region-level pairwise relationship between regions to calibrate the saliency values.

In general, current methods (*e.g.*, [15], [28], [29]) mainly adopt the binary cross-entropy loss function as the optimization objective, which is a pixel-level unary constraint for prediction by the formulation of Eq. (4). However, Eq. (4) only considers the relationship between each pixel and its corresponding ground-truth value, but does not take account of the relationship between different pixels or regions. In this manner, sometimes the saliency of whole local areas is completely detected incorrectly, which is caused by this problem lacking region-level relationship constraints.

To address this problem, we propose to model region-level pair-wise relationship as a supplement to the unary constraint and correct the probable errors. For the purpose of modeling the region-level relationship, we first construct a graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ for each image, where $\mathbb{V}$ and $\mathbb{E}$ are the sets of nodes and directed edges. In the graph, each node represents a region $v_i \in \mathbb{V}, i = 1, \ldots, N_v$ in images, where N_v is the number of regions in the image. Regions in an image are easily generated by some existing methods [55], [56] and we adopt SLIC algorithm [55] to over-segment an RGB image into super-pixels as regions with $N_v = 256$. And the edge $v_i \rightarrow v_j$ from the region v_i to the region v_j represents the relation between these two regions. We apply the locations of super-pixels of RGB images to its corresponding ground truths and predicted saliency maps, and then we get the regions of the ground truths and predicted saliency maps.

For the ground truth and predicted saliency map of an RGB image, we define the saliency value of a region as the average of the sum of the saliency values of each pixel in this region, and the saliency value of ith super-pixel is denoted as $\mathcal{S}_i (i = 1, \ldots, N_v)$. To model the relationship of regions, we use the subtraction of the saliency values between corresponding two nodes (*i.e.*, v_i and v_j) to represent the weight of each edge $v_i \rightarrow v_j$. Then, we construct structural matrix $\mathbf{M}$ to model the overall pair-wise relationship of an image as shown in Fig. 8. The value in ith row and jth column of $\mathbf{M}$ represents the weight of the edge $v_i \rightarrow v_j$. In this manner, we can construct structural matrices for the ground-truth mask as $\mathbf{M}_G$ and predicted saliency map as $\mathbf{M}_P$ of every image. The ground truth and saliency map are expected to have the similar structure, so we drive $\mathbf{M}_P$ to become the structural similarity matrix of $\mathbf{M}_G$ by minimizing the KL-divergence:

$$SS(P, G) = D_{KL}(N(\mathbf{M}_G)||N(\mathbf{M}_P)), \tag{14}$$

where $N(\cdot)$ is a normalization operation as used in Eq. (8), and P and G means the predicted saliency map and ground-truth mask of an image, respectively. This loss function is named as the structural similarity loss (denoted as SSL).

In this work, we conduct the SSL on the outputs $\phi^{(i)}_{\mathcal{RM}}(\pi^{(i)}_{\mathcal{RM}})$ of each stage in the purificatory network, and the formulation of the overall structural similarity loss is as follows:

$$L_{\mathcal{SS}} = \sum_{i=1}^{5} SS(\phi^{(i)}_{\mathcal{RM}}(\pi^{(i)}_{\mathcal{RM}}), G). \tag{15}$$

By taking the losses of Eqs. (3), (7), (8), (12) and (15), the overall learning objective can be formulated as follows:

$$\min_{\mathbb{P}} L_{\mathcal{P}} + L_{\mathcal{R},\mathcal{O}} + L_{\mathcal{R},\mathcal{E}} + L_{\mathcal{RM}} + L_{\mathcal{SS}}, \tag{16}$$

where $\mathbb{P}$ is the set of $\{\pi_i, \pi^{(i)}_{\mathcal{P}}, \pi^{(i)}_{\mathcal{R},\mathcal{O}}, \pi^{(i)}_{\mathcal{R},\mathcal{E}}, \pi^{(i)}_{\mathcal{RM}}\}_{i=1}^{5}$ for convenience of presentation.

V. Experiments

A. Experimental Setup

1) Datasets: To evaluate the performance of our method, we conduct experiments on six benchmark datasets [8], [23], [25]–[27]. Details of these datasets are briefly described as follows: ECSSD [8] consists of 1,000 images with complex and semantically meaningful objects. DUT-OMRON [23] has 5,168 complex images that are downsampled to a maximal side length of 400 pixels. PASCAL-S [24] includes 850 natural images that are pre-segmented into objects or regions with salient object annotation by eye-tracking test of 8 subjects. HKU-IS [25] contains 4,447 images which usually contain multiple disconnected salient objects or salient objects that touch image boundaries. DUTS [26] is a large scale dataset containing 10,533 training images (named as DUTS-TR) and 5019 test images(named as DUTS-TE). The images are challenging with salient objects that occupy various locations and scales as well as complex background. XPIE [27] is also a large dataset that has 10,000 images covering a variety of simple and complex scenes with various salient objects.

2) Evaluation Metrics: We choose mean absolute error (MAE), weighted F-measure score (F^w_β) [65], F-measure score (F_β), and F-measure curve to evaluate our method. MAE is the average pixel-wise absolute difference between ground-truth

TABLE I

PERFORMANCE ON SIX BENCHMARK DATASETS. SMALLER MAE, LARGER F_β^w AND F_β CORRESPOND TO BETTER PERFORMANCE. THE BEST RESULTS OF DIFFERENT BACKBONES ARE IN **BLUE** AND **RED** FONTS. "–" MEANS THE RESULTS CANNOT BE OBTAINED AND "†" MEANS THE RESULTS ARE POST-PROCESSED BY DENSE CONDITIONAL RANDOM FIELD (CRF) [57]. NOTE THAT THE BACKBONE OF PAGRN IS VGG-19 [44] AND THE ONE OF R3NET IS RESNEXT-101 [58]. MK: MSRA10K [43], DUTS: DUTS-TR [26], MB: MSRA-B [59]

Models	Training Dataset	ECSSD			DUT-OMRON			PASCAL-S			HKU-IS			DUTS-TE			XPIE		
		MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β
VGG-16 backbone [44]																			
KSR [60]	MB	0.132	0.633	0.810	0.131	0.486	0.625	0.157	0.569	0.773	0.120	0.586	0.773	-	-	-	-	-	-
HDHF [61]	MB	0.105	0.705	0.834	0.092	0.565	0.681	0.147	0.586	0.761	0.129	0.564	0.812	-	-	-	-	-	-
ELD [10]	MK	0.078	0.786	0.829	0.091	0.596	0.636	0.124	0.669	0.746	0.063	0.780	0.827	0.092	0.608	0.647	0.085	0.698	0.746
UCF [16]	MK	0.069	0.807	0.865	0.120	0.574	0.649	0.116	0.696	0.776	0.062	0.779	0.838	0.112	0.596	0.670	0.095	0.693	0.773
NLDF [13]	MB	0.063	0.839	0.892	0.080	0.634	0.715	0.101	0.737	0.806	0.048	0.838	0.884	0.065	0.710	0.762	0.068	0.762	0.825
Amulet [14]	MK	0.059	0.840	0.882	0.098	0.626	0.673	0.099	0.736	0.795	0.051	0.817	0.853	0.085	0.658	0.705	0.074	0.743	0.796
FSN [17]	MK	0.053	0.862	0.889	0.066	0.694	0.733	0.095	0.751	0.804	0.044	0.845	0.869	0.069	0.692	0.728	0.066	0.762	0.812
C2SNet [36]	MK	0.057	0.844	0.878	0.079	0.643	0.693	0.086	0.764	0.805	0.050	0.823	0.854	0.065	0.705	0.740	0.066	0.764	0.807
RA [29]	MB	0.056	0.857	0.901	0.062	0.695	0.736	0.105	0.734	0.811	0.045	0.843	0.881	0.059	0.740	0.772	0.067	0.776	0.836
PAGRN [34]	DUTS	0.061	0.834	0.912	0.071	0.622	0.740	0.094	0.733	0.831	0.048	0.820	0.896	0.055	0.724	0.804	-	-	-
RFCN [62]	MK	0.067	0.824	0.883	0.077	0.635	0.700	0.106	0.720	0.802	0.055	0.803	0.864	0.074	0.663	0.731	0.073	0.736	0.809
DSS† [63]	MB	0.052	0.872	0.918	0.063	0.697	**0.775**	0.098	0.756	0.833	0.040	0.867	**0.904**	0.056	0.755	0.810	0.065	0.784	0.849
MLM [9]	DUTS	0.045	0.871	0.897	0.064	0.681	0.719	0.077	0.778	0.813	0.039	0.859	0.882	0.049	0.761	0.776	-	-	-
AFNet [38]	DUTS	0.042	0.886	0.916	0.057	0.717	0.761	**0.073**	0.797	0.839	**0.036**	0.869	0.895	0.046	0.785	0.807	0.047	0.822	0.859
Ours	MK	0.042	0.887	0.914	0.064	0.708	0.743	0.080	0.779	0.830	0.042	0.852	0.885	0.052	0.768	0.792	0.053	0.808	0.851
Ours	DUTS	**0.040**	**0.892**	**0.920**	**0.054**	**0.730**	0.768	0.076	**0.798**	**0.841**	**0.036**	**0.870**	0.894	**0.043**	**0.797**	**0.816**	**0.044**	**0.830**	**0.868**
ResNet-50 backbone [53]																			
SRM [15]	DUTS	0.054	0.853	0.902	0.069	0.658	0.727	0.086	0.759	0.820	0.046	0.835	0.882	0.059	0.722	0.771	0.057	0.783	0.841
Picanet [28]	DUTS	0.047	0.866	0.902	0.065	0.695	0.736	0.077	0.778	0.826	0.043	0.840	0.878	0.051	0.755	0.778	0.052	0.799	0.843
R3† [64]	MK	0.040	0.902	0.924	0.063	0.728	0.768	0.095	0.760	0.834	0.036	0.877	0.902	0.057	0.765	0.805	0.058	0.805	0.854
DGRL [35]	DUTS	0.043	0.883	0.910	0.063	0.697	0.730	0.076	0.788	0.826	0.037	0.865	0.888	0.051	0.760	0.781	0.048	0.818	0.859
ICTBI [21]	DUTS	0.041	0.881	0.909	0.061	0.730	0.758	0.071	0.788	0.826	0.038	0.856	0.890	0.048	0.762	0.797	-	-	-
Ours	MK	0.038	0.896	0.917	0.063	0.728	0.754	0.075	0.801	0.842	0.036	0.871	0.892	0.047	0.780	0.802	0.046	0.823	0.860
Ours	DUTS	**0.035**	**0.907**	**0.928**	**0.051**	**0.747**	**0.776**	**0.070**	**0.805**	**0.847**	**0.031**	**0.889**	**0.904**	**0.039**	**0.817**	**0.829**	**0.041**	**0.843**	**0.876**

masks and estimated saliency maps. In computing F_β, we normalize the predicted saliency maps into the range [0, 255] and binarize the saliency maps with a threshold sliding from 0 to 255 to compare the binary maps with ground-truth masks. At each threshold, Precision and Recall can be computed. F_β is computed as:

$$F_\beta = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}, \tag{17}$$

where β^2 is set to 0.3 to emphasize more on Precision than Recall as suggested in [40]. Then we can plot F-measure curve based on all the binary maps over all saliency maps in a given dataset. We report F_β using an adaptive threshold for generating binary a saliency map and the threshold is computed as twice the mean of a saliency map. In addition, F_β^w is used to evaluate the overall performance (more details can be found in [65]).

3) Training and Inference: We train the networks in three stages and the training steps as follows: ① we first train the feature extractor and purificatory subnetwork with $L_{\mathcal{RM}}$ and $L_{\mathcal{SS}}$; ② we fix the purificatory subnetwork then train the promotion rectification subnetworks with $L_{\mathcal{P}}$, $L_{\mathcal{R},\mathcal{O}}$ and $L_{\mathcal{RE}}$; ③ Then we train the whole network with the overall loss in Eq. (16).

We use standard stochastic gradient descent algorithm to train our network end-to-end by optimizing the learning object in Eq. (16). In the optimization process, the parameters of feature extractor is initialized by the pre-trained backbone model [53], whose learning rate is set to 1×10^{-3} with a weight decay of 5×10^{-4} and momentum of 0.9. And the learning rate of rest layers are set to 10 times larger. Besides, we employ the "poly" learning rate policy for all experiments similar to [66].

We train our network with ResNet-50 [53] by utilizing the training set of DUTS-TR dataset [26] as used in [15], [28], [34], [35] and MSRA10K [64]. The training images are resized to the resolution of 320×320 for faster training, and applied horizontal flipping. For a more comprehensive demonstration and fairer comparison, we also use VGG-16 [44] as the backbone of our method instead of ResNet-50 [53], and train the new network without changing other settings. The training process takes about 20 hours and converges after 500k iterations (20k iterations for stage ①, 50k iterations for stage ② and 200k iterations for stage ③) with mini-batch of size 8 on a single NVIDA TITAN Xp GPU. During inference, the proposed network removes all the losses, and one image is directly fed into the network to produce the saliency map at the output of first stage in the purificatory network. And the network runs at about 27fps on a single NVIDIA 1080Ti GPU for inference.

B. Comparisons With the State-of-the-Art

We compare our approach denoted as with 19 state-of-the-art methods, including KSR [60], HDHF [61], ELD [10], UCF [16], NLDF [13], Amulet [14], FSN [17], SRM [15], C2SNet [36], RA [29], Picanet [28], PAGRN [34], R3Net [64], DGRL [35], RFCN [62], DSS [63], MLM [9], ICTBI [21] and AFNet [38]. We obtain the saliency maps of different methods from the authors or the deployment codes provided by the authors for fair comparison.

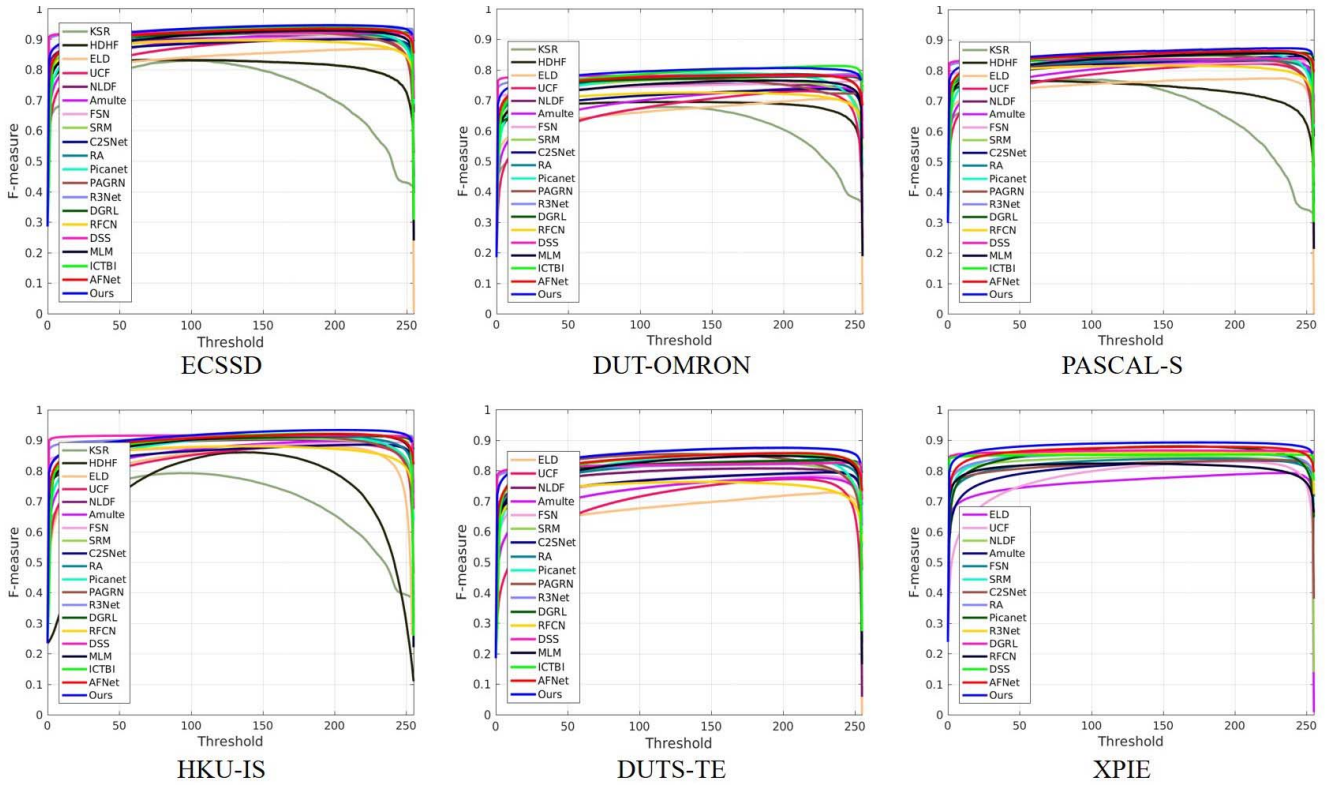

Fig. 9. The F-measure curves of 19 state-of-the-arts and our approach are listed across six benchmark datasets.

TABLE II

PERFORMANCE OF DIFFERENTS SETTING OF THE PROPOSED METHOD. PURNET IS THE PROPOSED METHOD. MEANINGS OF OTHER ABBREVIATIONS ARE AS FOLLOWS: RM: PURIFICATORY NETWORK, PA: PROMOTION ATTENTION NETWORK, RA: RECTIFICATION ATTENTION NETWORK, SSL: STRUCTURAL SIMILARITY LOSS

	RM	PA	RA	SSL	ECSSD			DUT-OMRON			PASCAL-S		
					MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β
Baseline	✓				0.046	0.864	0.895	0.064	0.700	0.725	0.077	0.776	0.820
Baseline + PA	✓	✓			0.044	0.877	0.919	0.055	0.719	0.760	0.079	0.778	0.838
Baseline + RA	✓		✓		0.043	0.878	0.917	0.053	0.725	0.765	0.076	0.781	0.838
Baseline + PM	✓	✓	✓		0.039	0.892	0.924	0.055	0.734	0.768	0.071	0.798	**0.848**
Baseline + SSL	✓			✓	0.043	0.880	0.917	0.057	0.716	0.760	0.074	0.786	0.838
PurNet	✓	✓	✓	✓	**0.035**	**0.907**	**0.928**	**0.051**	**0.747**	**0.776**	**0.070**	**0.805**	0.847

1) Quantitative Evaluation: We evaluate 19 state-of-the-art SOD methods and our method on six benchmark datasets with different backbones and training sets, and the results are listed in Tab. I. We can see that the proposed method consistently outperform other methods across all the six datasets, especially DUTS-TE and XPIE.

When training with ResNet-50, our method is noticeably improved from 0.765 to 0.817 on DUTS-TE and from 0.818 to 0.843 on XPIE compared to the second best results as for F_β^w. Also, it is worth noting that F_β of our method is significantly better compared with the second best results on DUTS-TE (0.829 against 0.805) and XPIE (0.876 against 0.859). As for MAE, our method has obvious advantages compared with other state-of-the-art algorithms on six datasets. Similarly, PurNet has an analogous and obvious improvement when training our network with VGG-16 as backbone. The advantages on these datasets confirm that our proposed purificatory mechanism and similarity structural loss can achieve great performance with different backbones.

For overall comparisons, F-measure curves of different methods are displayed in Fig. 9. We can observe that the F-measure curves of our approach are consistently higher than other state-of-the-art methods. These observations present the efficiency and robustness of our purificatory network across various challenging datasets, which indicates that the perspective of purificatory mechanism for the problem of SOD is useful. Note that the results of DSS, RA on HKU-IS [25] are only conducted on the test set.

Qualitative Evaluation: Some examples of saliency maps generated by our approach and other state-of-the-art algorithms are shown in Fig. 10. We can see that salient objects can pop-out with accurate location and details by the proposed method. From the row of 1 to 3 in Fig. 10, we can find that many methods usually can't locate the salient objects roughly. In our method, the salient objects are located with the help of effective promotion attention. In addition, lots of methods often mistakenly segment the details of salient objects. We think the reason for this error is that most existing

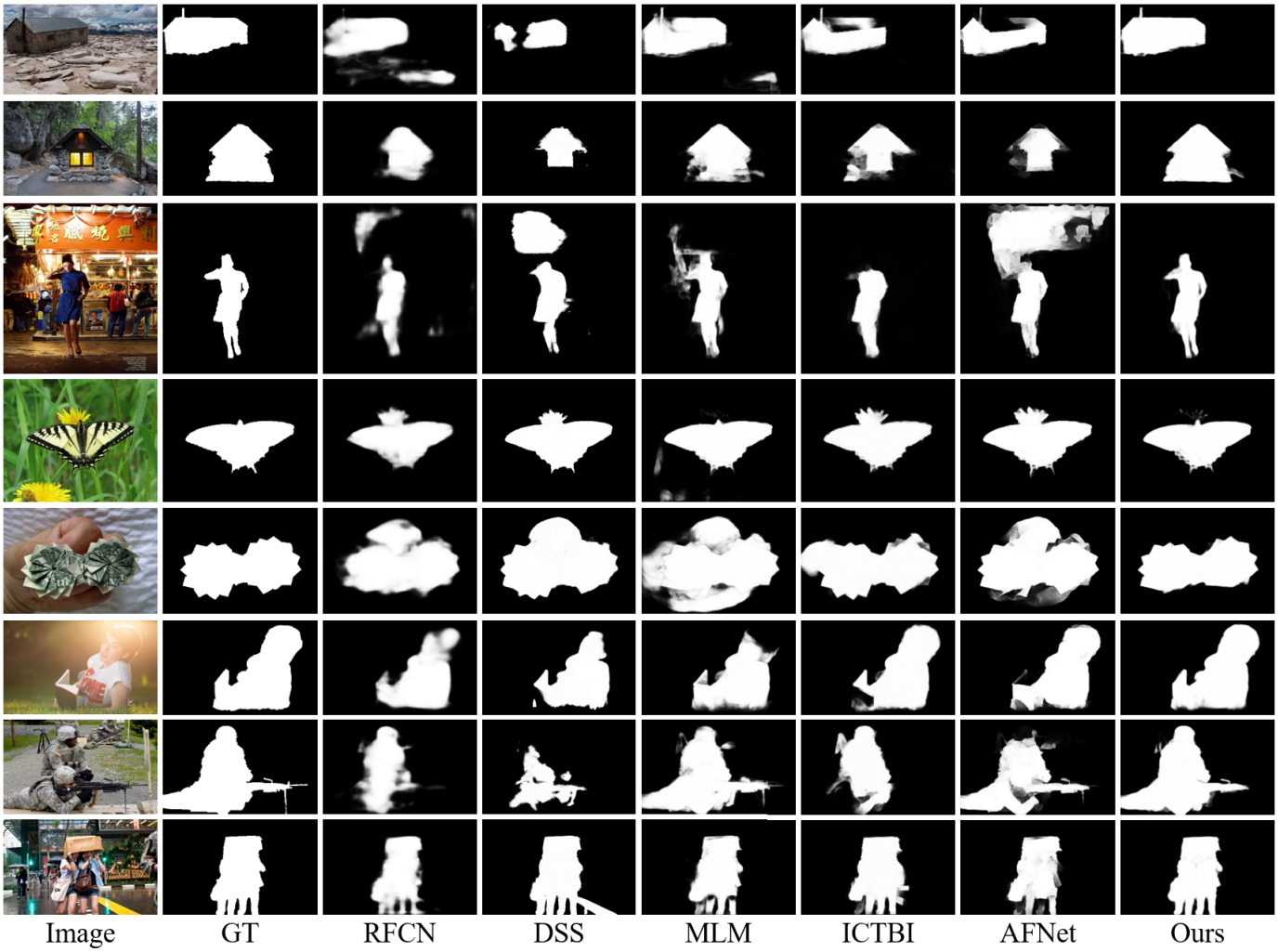

Fig. 10. Qualitative comparisons of the state-of-the-art algorithms and our approach. GT means ground-truth masks of salient objects.

TABLE III
COMPARISONS OF EACH SIDE-OUTPUTS AND THEIR FUSION. $Si(i = 1, \ldots, 5)$ MEANS iTH SIDE-OUTPUT AND FUSION MEANS THE AVERAGE OF S1 TO S5

	ECSSD			DUT-OMRON			PASCAL-S		
	MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β
S5	0.043	0.878	0.899	0.057	0.712	0.741	0.077	0.778	0.818
S4	0.043	0.880	0.899	0.057	0.715	0.742	0.076	0.781	0.819
S3	0.037	0.900	0.920	0.053	0.738	0.765	0.071	0.799	0.841
S2	**0.035**	0.906	0.927	**0.051**	0.746	0.775	**0.070**	0.804	0.846
S1	**0.035**	**0.907**	**0.928**	**0.051**	**0.747**	**0.776**	**0.070**	**0.805**	**0.847**
Fusion	0.038	0.894	0.918	0.054	0.733	0.758	0.073	0.794	0.837

methods usually lack the constraints of error-prone areas. From the row of 4 to 6 in Fig. 10, we can observe that our method achieves better performance, which indicates the ability of processing the fine structures and rectifying errors. More examples of complex scenes are shown in the row of 7 and 8, we can observe that the proposed method also obtains the impressive results. These observations indicate that addressing SOD from the perspective of purificatory mechanism and region-level pair-wise constraints is effective.

C. Ablation Studies

To validate the effectiveness of different components of the proposed method, we conduct several experiments on the benchmark datasets to compare the performance variations of our methods with different experimental settings.

1) Effectiveness of the Purificatory Mechanism: To investigate the effectiveness of the proposed purificatory mechanism, we conduct ablation experiments and introduce four different models for comparisons. The first setting is only the feature extractor and purificatory subnetwork, which is regarded as "Baseline". To explore the respective effectiveness of promotion attention and rectification attention, we conduct the second and third model by adding the promotion subnetwork (denoted as "Baseline + PA") and rectification subnetwork (denoted as "Baseline + RA"), respectively. In addition, we combine the two attention mechanisms (*i.e.*, purificatory mechanism) with the purificatory network as the fourth models, which is named as "Baseline + PM". We also list the proposed method with the purificatory mechanism and structural similarity loss as "**PurNet**".

The comparison results of above mentioned models are listed in Tab. II. We can observe that the promotion attention

TABLE IV

PERFORMANCE COMPARED WITH THE LATEST METHODS PUBLISHED IN 2019 AND 2020. SMALLER MAE, LARGER S-MEASURE AND E-MEASURE CORRESPOND TO BETTER PERFORMANCE. THE BEST AND SECOND RESULTS ARE IN **RED** AND **BLUE** FONTS

Models	Year	ECSSD			DUT-OMRON			HKU-IS			DUTS-TE		
		MAE	S-measure	E-measure	MAE	S-measure	E-measure	MAE	S-measure	E-measure	MAE	S-measure	E-measure
PAGENet [33]	2019	0.040	0.921	0.911	0.061	0.851	0.822	0.050	0.873	0.851	0.034	0.941	0.902
CPD-R [31]	2019	0.037	0.925	0.918	0.056	0.866	0.825	0.043	0.887	0.869	0.034	0.944	0.905
BASNet [39]	2019	0.037	0.921	0.916	0.056	**0.869**	0.836	0.048	0.884	0.866	0.032	0.946	0.909
ITSDNet [67]	2020	**0.034**	**0.927**	**0.925**	0.061	0.863	**0.840**	0.041	0.895	**0.885**	0.031	**0.952**	0.917
MINet [68]	2020	**0.033**	**0.927**	**0.925**	**0.055**	0.865	0.833	**0.037**	**0.898**	0.884	**0.029**	**0.953**	**0.919**
GCPANet [69]	2020	0.035	0.920	**0.927**	0.056	0.860	0.839	**0.038**	0.891	**0.891**	0.031	0.949	**0.920**
Ours	-	0.035	**0.925**	**0.925**	**0.051**	**0.868**	**0.841**	0.039	**0.897**	0.880	**0.031**	0.950	0.917

and rectification attention greatly improve the performance compared with "Baseline", which indicates the usefulness of the two attention mechanisms for SOD. In addition, we can find that "Baseline + RA" has better performance improvement than "Baseline + PA", which implies that the rectification of some error-prone areas is important to SOD. Moreover, a better performance has been achieved through the combination of the two attentions (*i.e.*, purificatory mechanism), which verifies the compatibility of the two attentions and effectiveness of the purificatory mechanism.

2) Effectiveness of the Structural Similarity Loss: To investigate the effectiveness of the proposed novel structural similarity loss (SSL), we conduct another experiments by only combining the loss with "Baseline" and this model is named as "Baseline + SSL". As listed in Tab. II, we can observe a remarkable improvement brought by SSL by comparing "Baseline" and "Baseline + SSL". The result shows that the loss plays an important role in the SOD task. In addition, by comparing "Baseline + PM" and "**PurNet**", we can find that SSL is still useful even when the results is advanced.

3) Performance of Each Side-Output: In order to explore how to obtain the best prediction of the proposed network, we conduct an additional experiment to compare the performance of each side-outputs and fusion in the purificatory subnetwork. As listed in Tab. III, we can see that the performance of last three side-outputs (*i.e.*, third, fourth and fifth side-output) is consistently worse than the one of the first two side-outputs (*i.e.*, first and second side-output). And the performance of fusion is lower than first, second and third side-output. The comparisons indicate the process of generating saliency maps in our network is progressively refined from the higher layer to the lower layer. Thus, we choose the first side-output as the results during inference.

4) Compared With the Latest Methods Published in 2019 and 2020 With New Evaluation Metrics: It is worth noting that since our method was submitted in 2019, many new works on salient object detection have been published since then. For a more fair comparison, we discussed and added several methods which are similar to our experimental settings for comparison. In addition, besides mean absolute error (MAE), we used two new evaluation metrics to verify the performance of the method from multiple perspectives, namely S-measure [70] and E-measure [71]. The experimental results are shown in Tab. IV. It can be seen that compared with the method published in 2019, our method has obvious advantages, and compared with the latest method published in 2020, our performance is still competitive.

VI. CONCLUSION

In this paper, we rethink the two difficulties that hinder the development of salient object detection. The difficulties consists of indistinguishable regions and complex structures. To solve these two issues, we propose the purificatory network with structural similarity loss. In this network, we introduce the promotion attention to improve the localization ability and semantic information for salient regions, which guides the network to focus on salient regions. We also propose the rectification subnetwork to provide the rectification attention for rectifying the errors. The two attentions are combined to form the purificatory mechanism to improve the promotable regions and rectifiable regions for purifying salient objects. Moreover, we also propose a novel region-level pair-wise structural similarity loss, which models and constrains the relationships between pair-wise regions. This loss can be used to be as a supplement to the unary constraint. Extensive experiments on six benchmark datasets have validated the effectiveness of the proposed approach.

REFERENCES

[1] W. James, F. Burkhardt, F. Bowers, and I. K. Skrupskelis, *The Principles of Psychology*, vol. 1, no. 2. London, U.K.: Macmillan, 1890.

[2] J. Li and W. Gao, *Visual Saliency Computation: A Machine Learning Perspective*, vol. 8408. Springer, 2014.

[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[4] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[5] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 5, pp. 769–779, May 2013.

[6] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 597–606.

[7] B. Lai and X. Gong, "Saliency guided dictionary learning for weakly-supervised image parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3630–3639.

[8] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1155–1162.

[9] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, and E. Ding, "A mutual learning method for salient object detection with intertwined multi-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8150–8159.

[10] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 660–668.

[11] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 678–686.

[12] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3203–3212.

[13] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6609–6617.

[14] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 202–211.

[15] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4019–4028.

[16] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 212–221.

[17] X. Chen, A. Zheng, J. Li, and F. Lu, "Look, perceive and segment: Finding the salient objects in images via two-stream fixation-semantic CNNs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1050–1058.

[18] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1741–1750.

[19] Y. Zeng, H. Lu, L. Zhang, M. Feng, and A. Borji, "Learning to promote saliency detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1644–1653.

[20] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient object detection driven by fixation prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1711–1720.

[21] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5968–5977.

[22] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3799–3808.

[23] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[24] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 280–287.

[25] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5455–5463.

[26] L. Wang *et al.*, "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 136–145.

[27] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4142–4150.

[28] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3089–3098.

[29] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 234–250.

[30] L. Zhang, J. Zhang, Z. Lin, H. Lu, and Y. He, "CapSal: Leveraging captioning to boost semantics for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6024–6033.

[31] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3907–3916.

[32] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3085–3094.

[33] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1448–1457.

[34] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 714–722.

[35] T. Wang *et al.*, "Detect globally, refine locally: A novel approach to saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3127–3135.

[36] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 355–370.

[37] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3917–3926.

[38] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1623–1632.

[39] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7479–7489.

[40] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

[41] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2334–2342.

[42] D. A. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2214–2219.

[43] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

[45] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[46] V. Mnih *et al.*, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 2204–2212.

[47] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5659–5667.

[48] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1831–1840.

[49] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[50] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[51] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," 2019, *arXiv:1904.09146*. [Online]. Available: http://arxiv.org/abs/1904.09146

[52] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[54] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–13.

[55] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[56] M. Van den Bergh, X. Boix, G. Roig, B. de Capitani, and L. Van Gool, "SEEDS: Superpixels extracted via energy-driven sampling," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2012, pp. 13–26.

[57] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2011, pp. 109–117.
[58] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.
[59] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.
[60] T. Wang, L. Zhang, H. Lu, C. Sun, and J. Qi, "Kernelized subspace ranking for saliency detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2016, pp. 450–466.
[61] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5012–5024, Nov. 2016.
[62] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Salient object detection with recurrent fully convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1734–1746, Jul. 2019.
[63] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.
[64] Z. Deng *et al.*, "R^3Net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 684–690.
[65] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 248–255.
[66] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," 2015, *arXiv:1506.04579*. [Online]. Available: http://arxiv.org/abs/1506.04579
[67] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9141–9150.
[68] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9413–9422.
[69] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," 2020, *arXiv:2003.00651*. [Online]. Available: http://arxiv.org/abs/2003.00651
[70] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.
[71] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," 2018, *arXiv:1805.10421*. [Online]. Available: http://arxiv.org/abs/1805.10421

Jia Li (Senior Member, IEEE) received the B.E. degree from Tsinghua University in 2005, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2011. He is currently a Full Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. Before he joined Beihang University in June 2014, he used to conduct research at Nanyang Technological University, Peking University, and the Shanda Innovations. He is the author or coauthor of over 90 technical articles in refereed journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), *International Journal of Computer Vision* (IJCV), IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), Conference on Computer Vision and Pattern Recognition (CVPR), and the International Conference on Computer Vision (ICCV). His research interests include computer vision and multimedia big data, especially the understanding, and generation of visual contents. He is a Senior Member of ACM, CIE, and CCF. He has been supported by the Research Funds for Excellent Young Researchers from the National Natural Science Foundation of China since 2019. In 2017, he was selected into the Beijing Nova Program and ever received the Second-Grade Science Award of Chinese Institute of Electronics in 2018. He received the two Excellent Doctoral Thesis Award from the Chinese Academy of Sciences in 2012 and the Beijing Municipal Education Commission in 2012. He received the First-Grade Science-Technology Progress Award from the Ministry of Education, China, in 2010.

Jinming Su received the B.S. degree from the School of Computer Science and Engineering, Northeastern University, in July 2017, and the master's degree with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, in January 2020. His research interests include computer vision, visual saliency analysis, and deep learning.

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.

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 computer vision, image salient object detection, and deep learning.

Yonghong Tian (Senior Member, IEEE) received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005. He is currently a Full Professor with the National Engineering Laboratory for Video Technology, School of Electronics Engineering and Computer Science, Peking University, Beijing. He has authored or coauthored more than 160 technical articles in refereed journals and conferences, and he has owned more than 57 Chinese and U.S. patents. His research interests include machine learning, computer vision, and multimedia big data. He is a Senior Member of CIE and CCF, and a member of ACM. He was a recipient of two national prizes and three ministerial prizes in China, and was a recipient of the 2015 EURASIP Best Paper Award for the *EURASIP Journal on Image and Video Processing*. He has served as the Technical Program Co-Chair for IEEE ICME 2015, IEEE BigMM 2015, IEEE ISM 2015, and IEEE MIPR 2018/2019, an organizing committee member of more than ten conferences, such as ACM Multimedia 2009, IEEE MMSP 2011, IEEE ISCAS 2013, IEEE ISM 2016, and BigMMs 2018, and a PC Member or the Area Chair of several conferences, such as CVPR, ICCV, KDD, AAAI, ACM MM, ECCV, and ICME. He is currently an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *IEEE Multimedia Magazine*, and IEEE ACCESS, and the Co-Editor-in-Chief of the *International Journal of Multimedia Data Engineering and Management*.