CLASS: Cross-Level Attention and Supervision for Salient Objects Detection

Lv Tang[1] and Bo Li*[2]

[1] State Key Lab for Novel Software Technology, Nanjing University, Nanjing, China
[2] Youtu Lab, Tencent, Shanghai, China
luckybird1994@gmail.com
libraboli@tencent.com

Abstract. Salient object detection (SOD) is a fundamental computer vision task. Recently, with the revival of deep neural networks, SOD has made great progresses. However, there still exist two thorny issues that cannot be well addressed by existing methods, indistinguishable regions and complex structures. To address these two issues, in this paper we propose a novel deep network for accurate SOD, named CLASS. First, in order to leverage the different advantages of low-level and high-level features, we propose a novel non-local cross-level attention (CLA), which can capture the long-range feature dependencies to enhance the distinction of complete salient object. Second, a novel cross-level supervision (CLS) is designed to learn complementary context for complex structures through pixel-level, region-level and object-level. Then the fine structures and boundaries of salient objects can be well restored. In experiments, with the proposed CLA and CLS, our CLASS net consistently outperforms 13 state-of-the-art methods on five datasets.

1 Introduction

Salient object detection (SOD) is a fundamental task in computer vision, which is derived with the goal of detecting and segmenting the most distinctive objects from visual scenes. As a preliminary step, SOD plays an essential role in various visual systems, such as object recognition [1, 2], semantic segmentation [3], visual tracking [4] and image-sentence matching [5].

Recently, with the application of deep convolutional neural networks (CNNs), salient object detection has achieved impressive improvements over conventional hand-crafted feature based approaches. Owing to their efficiency and powerful capability in visual feature representation, the CNN-based methods have pushed the performance of SOD to a new level, especially after the emergence of fully convolutional neural networks (FCNs). However, there still exist two thorny issues that cannot be well addressed by existing SOD methods. First, it is difficult to keep the uniformity and wholeness of the salient objects in some complex detecting scenes. As shown in Fig. 1(a)(b), some "salient-like" regions and large

* Correspondence should be addressed to Bo Li.

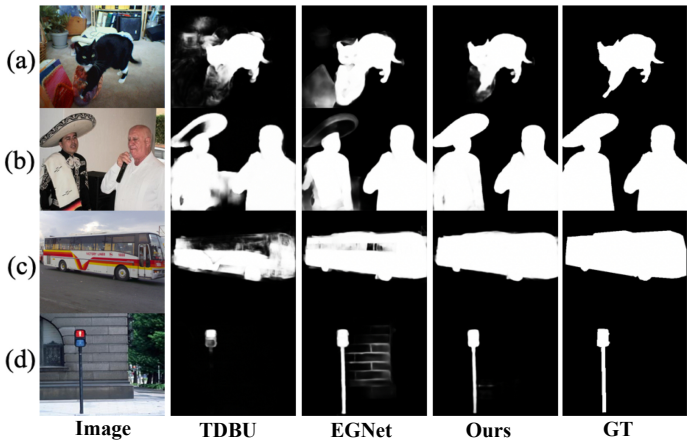

Fig. 1. Issues that cannot be well addressed by existing SOD methods. In (a)(b), some "salient-like" regions and large appearance change between salient object parts usually confuse the models to cause wrong predictions. In (c)(d), it is hard to maintain the fine structures and boundaries of salient objects. Images and ground-truth masks (GT) are from [6, 7]. Results are generated by TDBU [8], EGNet [9] and our approach.

appearance change between salient object parts usually confuse the models to cause wrong predictions. Second, it is hard to maintain the fine structures and boundaries of salient objects(see Fig. 1(c)(d)). These two issues hinder the further development of SOD, and make it still a challenging task.

To alleviate the first problem, some methods [10, 11, 8, 12–16] attempt to enhance the feature by aggregating multi-level and multi-scale features or adopting attention mechanisms to guide the models to focus on salient regions. However, these mechanisms ignore the relationships between the object parts and the complete salient object, leading to wrong prediction in complex real-world scenarios. For the second problem, methods [17, 18, 9, 19, 20] try to maintain the fine structures by introducing some special boundary branch or adding extra boundary supervision. These branches can provide boundary details to restore the salient contour, but they inevitably contain some noise edges might influence the final prediction(like the bricks in Fig.1(d)). Meanwhile, these pixel-level boundary supervisions not only cannot capture enough context of complex structures but need extra cost to get boundary labels.

In this paper, to address the aforementioned two issues, we propose a novel convolutional neural network, named CLASS, which achieves remarkable performance in detecting accurate salient objects. For the first issue, inspired by non-local mechanism [21, 22], we develop a novel attention module to capture the relationships between regions and the complete salient object. Unlike the conventional self-attention mechanism, we want to capture features dependencies through different levels, which is called cross-level attention module (CLA). On one hand, low-level features which contain the fine spatial details can guide

the selection of high-level through non-local position dependencies. Thus it can assist to locate preliminary salient objects and suppress the non-salient regions. On the other hand, high-level features with rich semantic information can be used as a guidance of low-level features through channel-wise dependencies, which can keep the wholeness of salient objects with large inner appearance change. For the second issue, in order to restore the fine structures of salient objects, we propose a novel cross-level supervision strategy (CLS). Unlike the pixel-level boundary loss, our CLS consists of binary cross entropy loss, a novel structural similarity loss and F-measure loss, which are designed to learn complementary information from ground truth through pixel-level, region-level and object-level. These cross-level constraints can provide context of complex structures to better calibrate the saliency values.

The main contributions of this paper can be summarize as:

(1) We propose a SOD network with a novel cross-level attention mechanism, which can keep the uniformity and wholeness of the detected salient objects by modeling the channel-wise and position-wise features dependencies through different levels.

(2) We introduce a novel cross-level supervision to train our network across three different levels: pixel-level, region-level and object-level. The complementarity between these losses can help restoring the fine structures and boundaries of salient objects.

(3) We conduct comprehensive experiments on five public SOD benchmark datasets. The results demonstrate that with the above two components the proposed CLASS net consistently outperforms state-of-the-art algorithms, which proves the effectiveness and superiority of our method.

2 Related Work

Over the past decades, a large amount of SOD algorithms have been developed. Traditional models [23–27] detect salient objects by utilizing various heuristic saliency priors with hand-crafted features. More details about the traditional methods can be found in the survey [28]. Here we mainly focus on deep learning based saliency detection models, especially the latest FCN-based methods in recent three years.

Lots of FCN-based models are devoted to exploring various feature enhancement strategies to improve the ability of localization and awareness of salient objects. Hou et al. [10] introduced short connections to the skip-layer structures within the HED [29] architecture, which provided rich multi-scale feature maps at each layer. Zhang et al. [12] aggregated multi-level feature maps into multiple resolutions, which were then fused to predict saliency maps in a recursive manner. Liu et al. [13] proposed a pixel-wise contextual attention to guide the network learning to attend global and local contexts. Chen et al. [15] propose a reverse attention network, which restore the missing object parts and details by erasing the current predicted salient regions from side-output features. Feng et al. [16] designed the attentive feedback modules to control the message passing

between encoder and decoder blocks. Wu et al. [11] introduced skip connection between multi-level features and a holistic attention module to refine the detection results by enlarging the coverage area of the initial saliency map. Wang et al. [8] proposed to integrate both top-down and bottom-up saliency inference by using multi-level features in an iterative and cooperative manner. However, these above mechanisms lack consideration of the relationships between the object parts and the complete salient object, leading to wrong prediction in complex real-world scenarios. Unlike these methods, we propose the cross-level attention module: the non-local position-wise and channel-wise features dependencies through different levels. The cross-level position attention can guide the network to suppress the non-salient regions, while the cross-level channel attention can keep the wholeness of salient objects with large inner appearance change.

Recently, some methods consider leveraging boundary information to restore the fine structures of salient objects. These methods usually utilize some special boundary branch or adding extra boundary supervision to get the boundary information. Li et al. [17] transferred salient knowledge from an existing contour detection model as useful priors to facilitate feature learning in SOD. In [18, 9, 20, 14], edge features from some sophisticated edge detection branches or modules were fused with salient features as complementary information to enhance the structural details for accurate saliency detection. However, these branches inevitably contain some noise edges that might influence the final prediction(like the bricks in Fig.1(d)). Liu et al. [19] proposed to utilize extra edge supervision to jointly train an edge detection branch and a SOD branch, which can assist the deep neural network to refine the details of salient objects. Feng et al. [16] presented a boundary-enhanced loss as a supplement to the cross-entropy loss for learning fine boundaries. These pixel-level boundary supervisions cannot capture enough context of complex structures and also increase labeling cost. Different form the above methods, our novel cross-level supervision strategy (CLS) , which consists of binary cross entropy loss, a novel structural similarity loss and F-measure loss, are designed to train our network across three different levels: pixel-level, region-level and object-level. With the learned complementary context of complex structures, it is much easier for our network to maintain the fine structures and boundaries of salient objects. For more information about the DNN-based methods, please refer to survey [30, 31].

3 Proposed Method

In this section, we first describe the overall architecture of the proposed deep salient object detection network, and then elaborate our main contributions, which are corresponding to cross-level attention module and cross-level supervision respectively.

3.1 Architecture

As illustrated in Fig. 2, the proposed CLASS net has a simple U-Net-like Encoder-Decoder architecture [32]. The ResNet-50 [33] is used as backbone feature en-

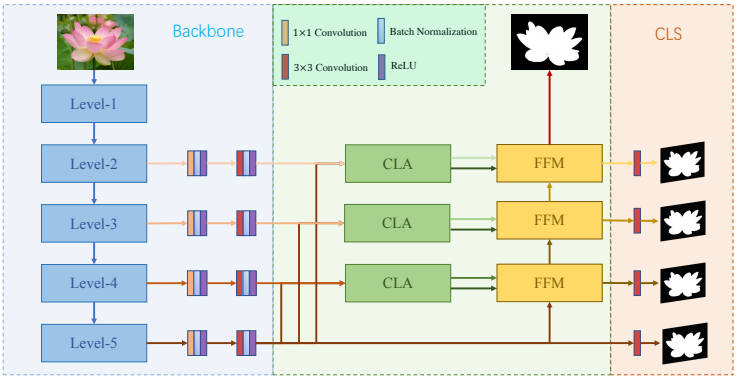

Fig. 2. An overview of proposed CLASS net. ResNet-50 is used as the backbone encoder. Cross-level attention module (CLA) is used to capture the long-range features dependencies between the high-level features and the low-level features. Feature fusion module (FFM) is a basic module to fuse features for decoder. Cross-level supervision (CLS) in each stages help to ease the optimization of CLASS net.

coder, which has five residual modules for encoding, named as level-1 to level-5 respectively. Because level-1 feature brings too much computational cost but little performance improvement, we don't use it for following process as suggested in work[11]. Between the encoder and decoder, we add two convolution blocks as the bridge. The 1×1 convolutional layer compresses the channels of high-level features for subsequent processing and the 3×3 convolutional layer transfers features for SOD task. Each of these convolution layers is followed by a batch normalization [34] and a ReLU activation [35]. The high-level feature in level-5 is denoted as $\{F_h|h = 5\}$, while the other three levels features are denoted as $\{F_l|l = 2, 3, 4\}$. Then cross-level attention modules are used to capture the long-range features dependencies between the high-level features (F_h) and the low-level features (F_l). For the decoder, we use a feature fusion module (FFM) to delicately aggregate the output features of CLA module in each stage and the upsampled features from the previous stage in a bottom-up manner. The output of each decoder stage is defined as $\{D_i|i = 2, 3, 4\}$. Cross-level supervision (CLS) is applied in each stage to train our CLASS net jointly. The output of the last stage is taken as the final saliency prediction.

3.2 Cross-Level Attention Module

Discriminant feature representations are essential for accurate SOD, while most existing methods cannot well keep the uniformity and wholeness of the salient objects in some complex scenes because of lacking consideration of the relationships between the indistinguishable regions and the salient object. To address this problem, inspired by non-local mechanism [21, 22], we develop a novel attention module to capture the long-range features dependencies. However, features

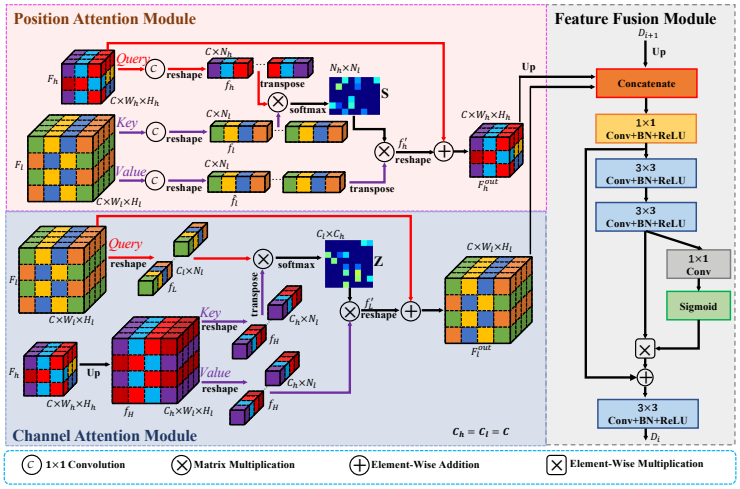

Fig. 3. An overview of the proposed Cross-Level Attention Module and Feature Fusion Module. Cross-Level Attention Module contains Position Attention and Channel Attention.

in different levels usually have different recognition information. Common non-local models [22], which rely on a single layer feature, exhibit limited ability in capturing sufficient long range dependencies. Unlike them, we want to leverage the advantages of features in different levels and propose the cross-level attention module. As illustrated in Fig. 3, we design two parts in CLA to model the channel-wise and position-wise features dependencies across the high-level feature and the low-levels features.

Position Attention Module. In some complex detecting scenes, there exist some non-salient regions which have "salient-like" appearance. These regions usually share some similar attributes with real salient regions like the high visual contrast. Thus, the saliency-like regions may also have high saliency semantics at the high-level layer. So the high-level feature which lacks low-level cues is difficult to distinguish saliency-like regions. We want to use the rich spatial details of low-level features as a guidance to make the high-level layer concentrate on real salient positions and then learn more discriminative features to suppress the non-salient regions. Specifically, as illustrated in Fig. 3, the input of Position Attention Module is a high-level feature map $F_h \in \mathcal{R}^{C\times H_h\times W_h}$ and a low-level feature map $F_l \in \mathcal{R}^{C\times H_l\times W_l}$. To be specific, for *Query* branch, we first add a 1×1 convolution layer on F_h and reshape the feature to $f_h \in \mathcal{R}^{C\times N_h}$, where $N_h = H_h \times W_h$. Meanwhile, for *Key* branch, we also use a 1×1 convolution layer on F_l and reshape the feature to $f_l \in \mathcal{R}^{C\times N_l}$, where $N_l = H_l \times W_l$. After that, we perform a matrix multiplication between the transpose of f_h and f_l, then apply a *softmax* function to calculate the spatial attention map $\mathbf{S} \in \mathcal{R}^{N_h\times N_l}$.

Each pixel value in $\mathbf{S}$ is defined as:

$$S(i,j) = \frac{exp(f_h^i \cdot f_l^j)}{\sum_{j=1}^{N_l} exp(f_h^i \cdot f_l^j)}, \tag{1}$$

where $i \in [1, N_h]$, $S(i,j)$ measures the j^{th} position in low-level feature impact on i^{th} position in high-level feature. Meanwhile, like *Key* branch, we generate feature $\hat{f}_l$ from *Value* branch and perform a matrix multiplication between S and the transpose of $\hat{f}_l$ to get $f'_h \in \mathcal{R}^{C \times N_h}$, which is defined as:

$$f'_h(i) = \sum_{j=1}^{N_l} S(i,j)\hat{f}_l(j), \tag{2}$$

Finally, we reshape f'_h to $\mathcal{R}^{C \times H_h \times W_h}$ and multiply it by a scale parameter α and perform an element-wise sum operation with F_h to obtain the final output $F_h^{out} \in \mathcal{R}^{C \times H_h \times W_h}$. It is defined as:

$$F_h^{out} = \alpha \cdot f'_h + F_h, \tag{3}$$

where α is initialized as 0 and gradually learns to assign more weight [36].

Channel Attention Module. In some complicated scenarios, salient objects may have large inner appearance change. These appearance variations are mainly reflected in the difference in the channels of low-level features. Since the channel of the low-level features contains almost no semantic information but low-level visual appearance cues, it is hard to maintain the semantic consistency of the object parts. To address this issue, we want to use the rich semantics of channels in high-level features to guide the selection of low-level features, which equips our network with the power of assigning saliency label to different-looking regions to keep the wholeness of salient objects. Specifically, as illustrated in Fig. 3, for channel attention module, we first use bilinear to upsample F_h to the spatial size of F_l, denoted as $f_H \in \mathcal{R}^{C_h \times H_l \times W_l}$, where $C_h = C$. For *Query* branch, we reshape F_l to $f_L \in \mathcal{R}^{C_l \times N_l}$, where $C_l = C$. For *Key* branch, we reshape f_H to $\mathcal{R}^{C_h \times N_l}$. Next, we perform a matrix multiplication between f_L and the transpose of f_H and apply a softmax function to get the channel attention map $\mathbf{Z} \in \mathcal{R}^{C_l \times C_h}$. Each pixel value in $\mathbf{Z}$ can be calculated as:

$$Z(i,j) = \frac{exp(f_L^i \cdot f_H^j)}{\sum_{j=1}^{C_h} exp(f_L^i \cdot f_H^j)}, \tag{4}$$

where $i \in [1, C_l]$, $Z(i,j)$ measures the j^{th} channel of high-level feature impact on i^{th} channel of low-level feature. At the same time, for *Value* branch, we reshape f_H to $\mathcal{R}^{C_h \times N_l}$ and perform a matrix multiplication with $\mathbf{Z}$ to get $f'_L \in \mathcal{R}^{C_l \times N_l}$, which is defined as:

$$f'_L(i) = \sum_{j=1}^{C_h} Z(i,j) f_H(j), \tag{5}$$

Finally, we reshape f'_L to $\mathcal{R}^{C_l \times H_l \times W_l}$ and multiply it by a scale parameter β and perform an element-wise sum operation with F_l to obtain the final output $F_l^{out} \in \mathcal{R}^{C \times H_l \times W_l}$. It is defined as:

$$F_l^{out} = \beta \cdot f'_L + F_l, \tag{6}$$

where β is initialized as 0 and gradually learns to assign more weight.

3.3 Feature Fusion Module

As illustrated in Fig.3, Each decoder network stage contains feature F_l^{out}, F_h^{out} from cross-level attention module, $D_{i+1} \in \mathcal{R}^{C \times \frac{H_l}{2} \times \frac{W_l}{2}}$ from previous decoder network stage. As these features contain different level information, we can not simply sum up these features for decoding. Inspired by SENet [37], we use an attention based feature fusion module to aggregate and refine these features effectively. Specifically, we first concatenate the three features then apply a 1×1 and two 3×3 convolutional layer with batch normalization and ReLU activation function to balance the scales of the features. Then we use a 1×1 convolutional layer and *sigmoid* function to compute a weight map, which amounts to feature selection and combination. Finally, guided by this weight map, we can archive an effective feature representation D_i for following process. Fig.3 shows the details of this design.

3.4 Cross-Level Supervision

Through the cross-level attention, features are enhanced for better keeping the uniformity and wholeness of the salient objects. Then we focus on restoring the fine structures and boundaries of salient objects. Toward this end, we propose a novel cross-level supervision strategy (CLS) to learn complementary context information from ground truth through pixel-level, region-level and object-level.

Let $\mathcal{I} = \{I_n\}_{n=1}^N$ and their groundtruth $\mathcal{G} = \{G_n\}_{n=1}^N$ denote a collection of training samples where N is the number of training images. After saliency detection, saliency maps are $\mathcal{S} = \{S_n\}_{n=1}^N$. In SOD, binary cross entropy (BCE) is the most widely used loss function, and it is a pixel-wise loss which is defined as:

$$L_{Pixel} = -\big(\ G_n log(S_n) + (1 - G_n) log(1 - S_n)\ \big). \tag{7}$$

From the formula of BCE loss, we find that it only considers the independent relationship between each pixel, which cannot capture enough context of complex structures, leading to blurry boundaries.

To address this problem, we propose to model region-level similarity as a supplement to the pixel-level constraint. Following the setting of [38, 39], we use the sliding window fashion to generate two corresponding regions from saliency map S_n and groundtruth G_n, denoted as $S_n^{region} = \{S_n^i : i = 1, ...M\}$ and $G_n^{region} = \{G_n^i : i = 1, ...M\}$, where M is the total number of region. Then, we adopt the simplified 2-Wasserstein distance[40, 41] to evaluate the distributional

similarity between S_n^i and G_n^i. Thus the proposed network can be trained by minimizing the similarity distance SSD_i between the corresponding regions, which is defined as:

$$SSD_i = ||\mu_{S_n^i} - \mu_{G_n^i}||_2^2 + ||\sigma_{S_n^i} - \sigma_{G_n^i}||_2^2, \tag{8}$$

where local statistics $\mu_{S_n^i}$, $\sigma_{S_n^i}$ is mean and std vector of S_n^i, $\mu_{G_n^i}$, $\sigma_{G_n^i}$ is mean and std vector of G_n^i. Finally, the overall loss function is defined as:

$$L_{Region} = \frac{1}{M}\sum_{i=1}^{M} SSD_i, \tag{9}$$

Pixel-level and region-level constraints can only capture local context for salient objects, a global constraint is still needed for accurate SOD. F-measure is often used to measure the overall similarity between the saliency map of the detected object and its groundtruth [42–44]. Hence we want to directly optimize the F-measure to learn the global information, called object-level supervision. For easy remembering, we denote F-measure as F_β in the following. The predicted saliency map S_n is a non-binary map, so we calculate F_β value via two steps. First, multiple thresholds are applied to the predicted saliency map to obtain multiple binary maps. Then, these binary maps are compared to the groundtruth. Hence, the whole process of calculating F_β is nondifferentiable. However, we can modify it to be differentiable. Considering pixel value $G_n(x, y)$ and $S_n(x, y)$, if $G_n(x, y) = 1$ and $S_n(x, y) = p$, it means this pixel has p probability to be true positive and $(1-p)$ probability to be false negative; if $G_n(x, y) = 0$ and $S_n(x, y) = p$, it means this pixel has p probability to be true negative and $1 - p$ to be false positive. So, we can calulate precision and recall by following Formulation:

$$precision = \frac{TP}{TP + FP} = \frac{S_n \cdot G_n}{S_n \cdot G_n + S_n \cdot (1 - G_n)} = \frac{S_n \cdot G_n}{S_n + \epsilon}, \tag{10}$$

$$recall = \frac{TP}{TP + FN} = \frac{S_n \cdot G_n}{S_n \cdot G_n + (1 - S_n) \cdot G_n} = \frac{S_n \cdot G_n}{G_n + \epsilon}, \tag{11}$$

$$F_\beta = \frac{(1 + \beta^2) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}, \tag{12}$$

where $\cdot$ means pixel-wise multiplication, $\epsilon = 1e^{-7}$ is a regularization constant to avoid division of zeros. L_{Object} loss function is defined as:

$$L_{Object} = 1 - F_\beta. \tag{13}$$

Note that all parts of our network are trained jointly, and the over all loss function is given as:

$$L = L_{Object} + L_{Region} + L_{Pixel}. \tag{14}$$

In addition, as show in Fig. 2, we use multi-level supervision as an as an auxiliary loss to facilitate sufficient training. The network has K levels and the whole loss is defined as:

$$L_{Final} = \sum_{i=1}^{K=4} \frac{1}{2^{i-1}} L_i. \tag{15}$$

In this loss function, high level loss has smaller weight because of its larger error. Finally, these cross-level constraints can provide context of complex structures to better calibrate the saliency values.

4 Experiments

4.1 Implementation Details

Following the works[18, 20, 9, 11], we train our proposed network on DUTS-TR. ResNet-50 [33] is used as the backbone network. For a more comprehensive demonstration, we also trained our network with VGG-16 [45] backbone. The whole network is trained end-to-end by stochastic gradient descent(SGD). Maximum learning rate is set to 0.005 for ResNet-50 or VGG-16 backbone and 0.05 for other parts. Warm-up and linear decay strategies are used to adjust the learning rate. Momentum and weight decay are set to 0.9 and 0.0005. Batchsize is set to 32 and maximum epoch is set to 100. We use Pytorch[3] to implement our model. Only horizontal flip and multi-scale input images are utilized for data augmentation as done in [18, 20, 15, 16]. A RTX 2080Ti GPU is used for acceleration. During testing, the proposed method runs at about 40 fps with about 352×352 resolution without any post-processing. Our code has been released [4].

We comprehensively evaluated our method on five representative datasets, including HKU-IS [46], ECSSD [47], PASCAL-S [6], DUT-OMRON [7] and DUTS [48], which contain 4447, 1000, 850, 5168 and 5019 images respectively. All datasets are human-labeled with pixel-wise ground-truth. Among them, more recent datasets PASCAL-S and DUT-TE are more challenging with salient objects that have large appearance change and complex background.

4.2 Evaluation Metrics

To evaluate the performance of the proposed method, four widely-used metrics are adopted: (1) Precision-Recall (PR) curve, which shows the tradeoff between precision and recall for different threshold (ranging from 0 to 255). (2) F-measure, (F_β), a weighted mean of average precision and average recall, calculated by $F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}$. We set β^2 to be 0.3 as suggested in [43]. (3) Mean Absolute Error (MAE), which characterize the average 1-norm distance between ground truth maps and predictions. (4) Structure Measure (S_m), a metric to

[3] https://pytorch.org/
[4] https://github.com/luckybird1994/classnet

evaluate the spatial structure similarities of saliency maps based on both region-aware structural similarity S_r and object-aware structural similarity S_o, defined as $S_\alpha = \alpha * S_r + (1 - \alpha) * S_o$, where $\alpha = 0.5$ [39].

Table 1. Performance of 13 sotas and the proposed method on five benchmark datasets. Smaller MAE, larger F_β and S_m correspond to better performance. The best results of different backbones are in blue and red fonts. "†" means the results are post-processed by dense conditional random field(CRF) [49]. MK: MSRA10K [24], DUTS: DUTS-TR [48], MB: MSRA-B [50].

Models	Training dataset	ECSSD			DUTS-TE			DUT-OMRON			PASCAL-S			HKU-IS		
		F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE
VGG-16 backbone																
Amulet(ICCV2017) [12]	MK	0.868	0.894	0.059	0.678	0.804	0.085	0.647	0.781	0.098	0.757	0.814	0.097	0.841	0.886	0.051
C2SNet(ECCV2018) [17]	MK	0.853	0.882	0.059	0.710	0.817	0.066	0.664	0.780	0.079	0.754	0.821	0.085	0.839	0.873	0.051
RAS(ECCV2018) [15]	MB	0.889	0.893	0.056	0.751	0.839	0.059	0.713	0.814	0.062	0.777	0.792	0.101	0.871	0.887	0.045
PiCA-V(CVPR2018) [13]	DUTS	0.885	0.914	0.046	0.749	0.861	0.054	0.710	0.826	0.068	0.789	0.842	0.077	0.870	0.906	0.042
DSS†(TPAMI2019) [10]	MB	0.904	0.882	0.052	0.808	0.820	0.057	0.740	0.790	0.063	0.801	0.792	0.093	0.902	0.878	0.040
PAGE(CVPR2019) [14]	MK	0.906	0.912	0.042	0.777	0.854	0.052	0.736	0.824	0.062	0.806	0.835	0.075	0.882	0.903	0.037
AFNet(CVPR2019) [16]	DUTS	0.908	0.913	0.042	0.792	0.867	0.046	0.738	0.826	0.057	0.820	0.848	0.070	0.888	0.905	0.036
CPD-V(CVPR2019) [11]	DUTS	0.915	0.910	0.040	0.813	0.867	0.043	0.745	0.818	0.057	0.820	0.838	0.072	0.896	0.904	0.033
TSPOA(ICCV2019) [51]	DUTS	0.900	0.907	0.046	0.776	0.860	0.049	0.716	0.818	0.061	0.803	0.836	0.076	0.882	0.902	0.038
BANet-V(ICCV2019) [18]	DUTS	0.910	0.913	0.041	0.789	0.861	0.046	0.731	0.819	0.061	0.812	0.834	0.078	0.887	0.902	0.037
EGNet-V(ICCV2019) [9]	DUTS	0.913	0.913	0.041	0.800	0.878	0.044	0.744	0.813	0.057	0.809	0.837	0.076	0.893	0.910	0.035
Ours	DUTS	0.917	0.915	0.038	0.833	0.880	0.039	0.749	0.820	0.057	0.838	0.853	0.062	0.909	0.915	0.031
ResNet50 backbone																
PiCA-R(CVPR2018) [13]	DUTS	0.886	0.917	0.046	0.759	0.869	0.051	0.717	0.832	0.065	0.792	0.848	0.074	0.870	0.904	0.043
TDBU(CVPR2019) [8]	MK	0.880	0.918	0.041	0.767	0.865	0.048	0.739	0.837	0.061	0.775	0.844	0.070	0.878	0.907	0.038
CPD-R(CVPR2019) [11]	DUTS	0.917	0.918	0.037	0.805	0.869	0.043	0.747	0.825	0.056	0.820	0.842	0.070	0.891	0.905	0.034
SCRN(ICCV2019) [20]	DUTS	0.918	0.927	0.037	0.808	0.885	0.040	0.746	0.837	0.056	0.827	0.848	0.062	0.896	0.916	0.034
BANet(ICCV2019) [18]	DUTS	0.923	0.924	0.035	0.815	0.879	0.040	0.746	0.832	0.059	0.823	0.845	0.069	0.900	0.913	0.032
EGNet(ICCV2019) [9]	DUTS	0.920	0.925	0.037	0.815	0.887	0.039	0.755	0.837	0.053	0.817	0.846	0.073	0.901	0.918	0.031
Ours	DUTS	0.933	0.928	0.033	0.856	0.894	0.034	0.774	0.838	0.052	0.849	0.863	0.059	0.921	0.923	0.028

4.3 Comparisons with the State-of-the-Arts

We compare our approach CLASS net with 13 state-of-the-art methods, including Amulet [12], C2SNet [17], RAS [15], PiCAnet [13], DSS [10], PAGE [14], AFNet [16], CPD [11], TSPOANet [51], TDBU [8], SCRN [20], BANet [18] and EGNet [9]. For fair comparison, we obtain the saliency maps of these methods from authors or the deployment codes provided by authors.

Quantitative Evaluation. The proposed approach is compared with 13 state-of-the-art SOD methods on five datasets, and the results are reported in Table 1 and Fig. 4. From Table 1, we can see that our method consistently outperforms other methods across all the five benchmark datasets. It is noteworthy that our method improves the F-measure and S-measure achieved by the best-performing existing algorithms by a large margin on two challenging datasets PASCAL-S (F_β: 0.849 against 0.827, S_m: 0.863 against 0.848) and DUTS-TE (F_β: 0.856 against 0.815, S_m: 0.894 against 0.887). As for MAE, our method obviously exceed other state-of-the-art algorithms on all five datasets. When using VGG-16 as backbone, our method still consistently outperfrom other methods, which verifies that our proposed CLA and CLS can achieve great performance with different backbone. For overall comparisons, PR curves of different methods are displayed in Fig. 4. One can observe that our approach noticeably higher

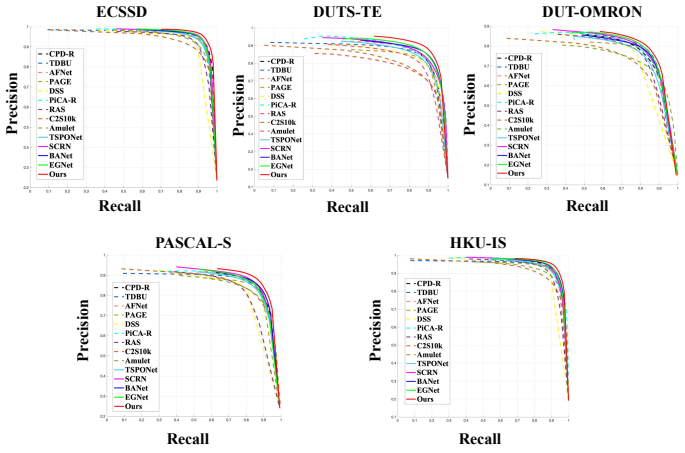

Fig. 4. Comparison of the PR curves across five benchmark datasets.

than all the other methods. These observations present the efficiency and robustness of our CLASS net across various challenging datasets, which indicates that the perspective of CLA for the problem of SOD is useful.

Qualitative Evaluation. To exhibit the superiority of the proposed approach, Fig. 5 show representative examples of saliency maps generated by our approach and other state-of-the-art algorithms. As can be seen, the proposed method can keep the uniformity and wholeness of the salient objects meanwhile maintain the fine structures and boundaries in various challenging scenes. From the column of 1 and 2 in Fig. 5, we can observe that with the influence of "salient-like" regions (mountain and water reflection), existing methods usually give wrong predictions. While, in our method, by the guidance of position-wise cross-level attention, the salient objects are accurately located and the non-salient regions are well suppressed. Example in third column with large inner appearance change can cause incomplete detection problem in existing methods. With the help of channel-wise cross-level attention, our method can better keep the wholeness of the salient object. Moreover, for the case of multiple and small objects in the of 4 to 6, our method can detect all the salient objects with the relationship information captured by cross-level attention, whereas the other methods mostly miss objects or introduce some background noise. From the column of 7 and 8, we can find that most existing methods cannot maintain the fine structures and boundaries of objects in the case of low contrast between salient object and background as well as the complicated scene. Note that some methods with special edge branches (EGNet, BANet and SCRN) can keep some structural details of example in column 8 and 9. However, These branches inevitably contain some noise edges can introduce background noise in the final prediction. It can be clearly observed that our method achieves impressive

performance in all these cases, which indicates the effectiveness of cross-level supervision in maintaining the fine structures and boundaries of salient objects.

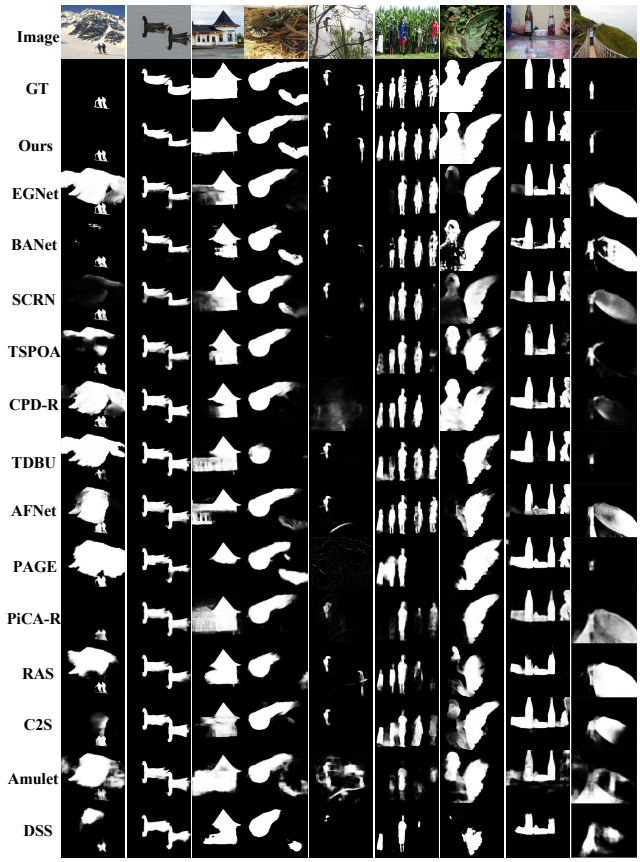

Fig. 5. Qualitative comparisons of the state-of-the-art algorithms and our approach.

4.4 Ablation Studies

To validate the effectiveness of the proposed components of our method, we conduct a series of experiments on three datasets with different settings.

Supervision ablation. To investigate the effectiveness of our proposed cross-level supervision, we conduct a set of experiments over different losses based on a baseline U-Net architecture. As listed in Table 2, we can observe a remarkable and consistent improvement brought by different level supervisions. Compared with only using pixel-level supervision, adding region-level structural similarity supervision can significantly improve the performance on all three

Table 2. Ablation study on different settings of supervision and architecture.

Ablation		Configurations							ECSSD			PASCAL-S			DUT-TE		
		Pixel Level	Region Level	Object Level	MS	CLA-C	CLA-P	FFM	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE
Loss	1.	√							0.900	0.910	0.043	0.808	0.838	0.072	0.782	0.873	0.044
	2.	√	√						0.915	0.918	0.040	0.825	0.846	0.068	0.820	0.882	0.039
	3.	√	√	√					0.918	0.919	0.039	0.832	0.850	0.067	0.837	0.885	0.038
	4.	√	√	√	√				0.920	0.920	0.038	0.838	0.853	0.064	0.841	0.887	0.037
Architecture	5.	√	√	√	√			√	0.923	0.922	0.037	0.842	0.855	0.062	0.845	0.888	0.036
	6.	√	√	√	√	√	√		0.930	0.926	0.034	0.847	0.860	0.060	0.852	0.892	0.035
	7.	√	√	√	√	√		√	0.927	0.924	0.036	0.845	0.859	0.062	0.850	0.889	0.036
	8.	√	√	√	√		√	√	0.926	0.925	0.036	0.844	0.858	0.061	0.851	0.891	0.035
	9.	√			√	√	√	√	0.919	0.925	0.037	0.830	0.861	0.063	0.813	0.891	0.039
	10.	√	√	√	√	√	√	√	0.933	0.928	0.033	0.849	0.863	0.059	0.856	0.894	0.034

metrics, especially the S-measure, which shows its ability of maintaining fine structures and boundaries of salient objects. Object-level supervision further improve the performance on F-measure. When these supervision are combined and applied at each stage (MS), we can get the best SOD results. In addition, by comparing setting No.9 and No.10, we can find that CLS is still useful even when the results is advanced.

Architecture ablation. To prove the effectiveness of our CLASS net, we report the quantitative comparison results of our model with different architectures. As shown in Table 2, Comparing No.5 and No.4, only using FFM can slightly improve the performance. Comparing No.6 and No.4, More significant improvements can be observed when we add channel-wise cross-level attention(CLA-C) and position-wise cross-level attention(CLA-P). Comparing No.7 with No.5, or No.8 with No.5, independently using CLA-C or CLA-P can also improve the performance. Finally, a best performance can be achieved through the combination of the CLA and FFM compared with baseline architecture(No.4), which verifies the compatibility of the two attentions and effectiveness of the features fusion module. For more comprehensive analyses of our proposed method, please refer to the supplementary materials[5].

5 Conclusions

In this paper, we revisit the two thorny issues that hinder the development of salient object detection. The issues consist of indistinguishable regions and complex structures. To address these two issues, in this paper we propose a novel deep network for accurate SOD, named CLASS. For the first issue, we propose a novel non-local cross-level attention (CLA), which can leverage the advantages of features in different levels to capture the long-range feature dependencies. With the guidance of the relationships between low-level and high-level features, our model can better keep the uniformity and wholeness of the salient objects in some complex scenes. For the second issue, a novel cross-level supervision (CLS) is designed to learn complementary context for complex structures through pixel-level, region-level and object-level. Then the fine structures and boundaries of salient objects can be well restored. Extensive experiments on five benchmark datasets have validated the effectiveness of the proposed approach.

[5] https://arxiv.org/abs/2009.10916

References

1. Rutishauser, U., Walther, D., Koch, C., Perona, P.: Is bottom-up attention useful for object recognition? In: CVPR (2). (2004) 37–44(2014)
2. Ren, Z., Gao, S., Chia, L., Tsang, I.W.: Region-based saliency detection and its application in object recognition. IEEE Trans. Circuits Syst. Video Techn. **24** (2014) 769–779
3. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M., Feng, J., Zhao, Y., Yan, S.: STC: A simple to complex framework for weakly-supervised semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39** (2017) 2314–2320
4. Hong, S., You, T., Kwak, S., Han, B.: Online tracking by learning discriminative saliency map with convolutional neural network. In: ICML(2015). Volume 37 of JMLR Workshop and Conference Proceedings., JMLR.org (2015) 597–606(2015)
5. Ji, Z., Wang, H., Han, J., Pang, Y.: Saliency-guided attention network for image-sentence matching. In: ICCV(2019). (2019)
6. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: CVPR. (2014) 280–287(2014)
7. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.: Saliency detection via graph-based manifold ranking. In: CVPR. (2013) 3166–3173
8. Wang, W., Shen, J., Cheng, M., Shao, L.: An iterative and cooperative top-down and bottom-up inference network for salient object detection. In: CVPR, Computer Vision Foundation / IEEE (2019) 5968–5977(2019)
9. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: ICCV(2019). (2019)
10. Hou, Q., Cheng, M., Hu, X., Borji, A., Tu, Z., Torr, P.H.S.: Deeply supervised salient object detection with short connections. IEEE Trans. Pattern Anal. Mach. Intell. **41** (2019) 815–828
11. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: CVPR. (2019) 3907–3916(2019)
12. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: ICCV. (2017) 202–211
13. Liu, N., Han, J., Yang, M.: Picanet: Learning pixel-wise contextual attention for saliency detection. In: CVPR. (2018) 3089–3098(2018)
14. Wang, W., Zhao, S., Shen, J., Hoi, S.C.H., Borji, A.: Salient object detection with pyramid attention and salient edges. In: CVPR. (2019) 1448–1457(2019)
15. Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection. In: ECCV(2018). Volume 11213., Springer (2018) 236–252
16. Feng, M., Lu, H., Ding, E.: Attentive feedback network for boundary-aware salient object detection. In: CVPR. (2019) 1623–1632(2019)
17. Li, X., Yang, F., Cheng, H., Liu, W., Shen, D.: Contour knowledge transfer for salient object detection. In: ECCV (2018). Volume 11219., Springer (2018) 370–385
18. Su, J., Li, J., Zhang, Y., Xia, C., Tian, Y.: Selectivity or invariance: Boundary-aware salient object detection. In: ICCV(2019). (2019)
19. Liu, J., Hou, Q., Cheng, M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: CVPR. (2019) 3917–3926(2019)
20. Wu, Z., Su, L., Huang, Q.: Stacked cross refinement network for edge-aware salient object detection. In: ICCV(2019). (2019)
21. Buades, A., Coll, B., Morel, J.: A non-local algorithm for image denoising. In: CVPR. (2005) 60–65(2005)

22. Wang, X., Girshick, R.B., Gupta, A., He, K.: Non-local neural networks. In: CVPR. (2018) 7794–7803(2018)
23. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20** (1998) 1254–1259
24. Cheng, M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.: Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell. **37** (2015) 569–582
25. Wang, J., Jiang, H., Yuan, Z., Cheng, M., Hu, X., Zheng, N.: Salient object detection: A discriminative regional feature integration approach. International Journal of Computer Vision **123** (2017) 251–268
26. Wang, T., Zhang, L., Lu, H., Sun, C., Qi, J.: Kernelized subspace ranking for saliency detection. In: ECCV(2016). Volume 9912. (2016) 450–466
27. Klein, D.A., Frintrop, S.: Center-surround divergence of feature statistics for salient object detection. In: ICCV. (2011) 2214–2219(2011)
28. Borji, A., Cheng, M., Hou, Q., Jiang, H., Li, J.: Salient object detection: A survey. Computational Visual Media **5** (2019) 117–150
29. Xie, S., Tu, Z.: Holistically-nested edge detection. International Journal of Computer Vision **125** (2017) 3–18
30. Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H.: Salient object detection in the deep learning era: An in-depth survey. CoRR **abs/1904.09146** (2019)
31. Han, J., Zhang, D., Cheng, G., Liu, N., Xu, D.: Advanced deep-learning techniques for salient and category-specific object detection: A survey. IEEE Signal Process. Mag. **35** (2018) 84–100
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015). Volume 9351. (2015) 234–241
33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016) 770–778(2016)
34. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML(2015). Volume 37., JMLR.org (2015) 448–456
35. Hahnloser, R.H.R., Seung, H.S.: Permitted and forbidden sets in symmetric threshold-linear networks. In: NIPS, MIT Press (2000) 217–223
36. Zhang, H., Goodfellow, I.J., Metaxas, D.N., Odena, A.: Self-attention generative adversarial networks. In: ICML(2019). Volume 97. (2019) 7354–7363
37. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. (2018) 7132–7141(2018)
38. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Processing **13** (2004) 600–612
39. Fan, D., Cheng, M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: ICCV. (2017) 4558–4567(2017)
40. Berthelot, D., Schumm, T., Metz, L.: BEGAN: boundary equilibrium generative adversarial networks. CoRR **abs/1703.10717** (2017)
41. He, R., Wu, X., Sun, Z., Tan, T.: Wasserstein CNN: learning invariant features for NIR-VIS face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **41** (2019) 1761–1773
42. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps. In: CVPR. (2014) 248–255(2014)
43. Borji, A., Cheng, M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE Trans. Image Processing **24** (2015) 5706–5722

44. Wang, W., Shen, J., Dong, X., Borji, A.: Salient object detection driven by fixation prediction. In: CVPR, IEEE Computer Society (2018) 1711–1720
45. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. (2015)
46. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: CVPR. (2015) 5455–5463(2015)
47. Shi, J., Yan, Q., Xu, L., Jia, J.: Hierarchical image saliency detection on extended CSSD. Trans. Pattern Anal. Mach. Intell. **38** (2016) 717–729
48. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: CVPR. (2017) 3796–3805(2017)
49. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS. (2011) 109–117
50. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.: Learning to detect a salient object. IEEE Trans. Pattern Anal. Mach. Intell. **33** (2011) 353–367
51. Liu, Y., Zhang, Q., Zhang, D., Han, J.: Employing deep part-object relationships for salient object detection. In: ICCV(2019). (2019)