# Cross-layer Feature Pyramid Network for Salient Object Detection

Zun Li[1], Congyan Lang[1], Jun Hao Liew[2], Yidong Li[1], Qibin Hou[2], Jiashi Feng[2]
[1]Beijing Jiaotong University, [2]National University of Singapore
lznus2018@gmail.com, cylang@bjtu.edu.cn, liewjunhao@u.nus.edu, ydli@bjtu.edu.cn, andrewhoux@gmail.com, elefjia@nus.edu.sg

## Abstract

*Feature pyramid network (FPN) based models, which fuse the semantics and salient details in a progressive manner, have been proven highly effective in salient object detection. However, it is observed that these models often generate saliency maps with incomplete object structures or unclear object boundaries, due to the* indirect *information propagation among distant layers that makes such fusion structure less effective. In this work, we propose a novel Cross-layer Feature Pyramid Network (CFPN), in which direct cross-layer communication is enabled to improve the progressive fusion in salient object detection. Specifically, the proposed network first aggregates multi-scale features from different layers into feature maps that have access to both the high- and low-level information. Then, it distributes the aggregated features to all the involved layers to gain access to richer context. In this way, the distributed features per layer own both semantics and salient details from all other layers simultaneously, and suffer reduced loss of important information. Extensive experimental results over six widely used salient object detection benchmarks and with three popular backbones clearly demonstrate that CFPN can accurately locate fairly complete salient regions and effectively segment the object boundaries.*

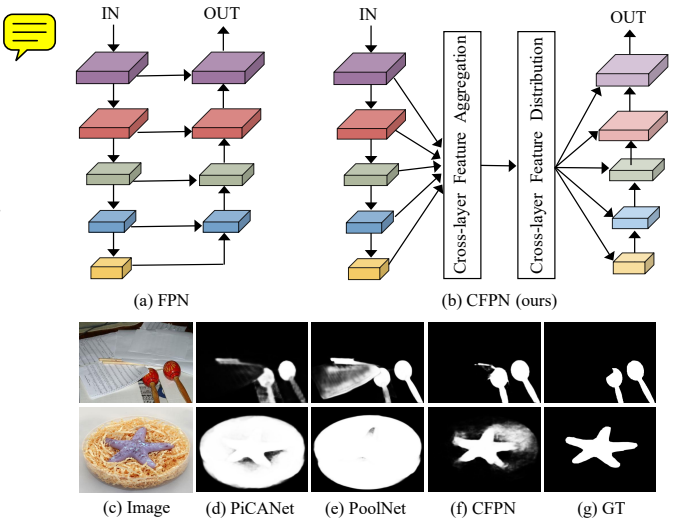



Figure 1: Illustration of existing feature pyramid fusion based structure and the proposed CFPN. Top panel: (a) existing FPN based context fusion structure; (b) pipeline of the proposed cross-layer feature pyramid network (CFPN). Bottom Panel: (d) and (e) are examples of saliency maps produced by vanilla FPN based saliency methods PiCANet [21], PoolNet [19]; (f) saliency maps generated by our CFPN. Clearly, saliency maps produced by CFPN show clearer object contour and look closer to the ground truth.

## 1. Introduction

Salient object detection aims to locate and segment the most visually distinctive objects or regions in a given image. It serves as a fundamental step in many computer vision tasks like object segmentation [38, 40], visual tracking [8, 10] and photo cropping [39]. Recently, deep learning based approaches [20, 35, 11, 3, 47, 34, 21, 19, 42, 27] have achieved remarkable performance in salient object detection, outperforming the traditional methods [30, 4, 45, 44] by a large margin. Among them, those leveraging pyramid style fusion [18, 28, 26, 41], especially the feature pyramid network (FPN [18]) that progressively fuses multi-scale features in a top-down pathway, have received great attention due to their effectiveness for improving localization accuracy and recovering boundary details.

Despite their good performance, there is still a large room of improvement for this fusion based approach. As shown in Fig. 1 (a), the pyramid fusion structure stage-wisely fuses high-level semantics with low-level details via lateral connections. However, two drawbacks exist in such approach. First, the low-level visual information, such as object edge can only be accessed at the final fusion stage, making predicted saliency maps from those methods have low-quality object boundaries. Second, as pointed out in [19, 42], in such pyramid fusion structure, the high-level semantics are progressively transmitted to the shallower layers, and hence the semantically salient cues captured by deeper layers may be gradually diluted throughout the progressive fusion. As a result, the predicted results tend

to have incomplete object structures or over-predicted foreground regions. To alleviate this limitation, attention models [21, 7, 37], gate functions [1, 51, 46], multi-scale feature integration [34, 49], and extra supervision (e.g., edge detection [19], boundary loss [23]) have been proposed in the literature. However, the information propagation is mainly limited between adjacent layers[1] at each fusion stage. Thus, these models still suffer from similar problems as illustrated in Fig. 1 (d)-(e).

In this work, we propose a novel Cross-layer Feature Pyramid Network (CFPN) aiming at directly exchanging information across different layers and further boosting information propagation for better salient object detection. As illustrated in Fig. 1 (b), CFPN is built on FPN but adopts the following novel architecture designs. First, it contains a Cross-layer Feature Aggregation module (CFA) that incorporates multi-scale features from different and distant layers to allow communication among different layers. Among which, CFA dynamically generates a set of layer-specific aggregation weights to weigh different layer features according to their usefulness for salient object detection. Second, given the reweighted features from CFA, CFPN also contains a Cross-layer Feature Distribution module (CFD) to allocate the aggregated features to their corresponding layers for the subsequent stage-wise fusion. Collaborating with CFD, the distributed features at each layer have access to both semantics and fine details from all other layers simultaneously, and hence reducing the loss of important information during the progressive fusion. As a result, better saliency maps can be obtained as shown in Fig. 1 (f). Clearly, benefiting from more direct information propagation among all the layers, CFPN can predict more complete salient objects with more accurate boundaries.

Our main contributions are summarized as follows:

- Through analyzing performance limitation of FPN-alike models, we propose that establishing direct information communication across multiple layers is important for salient object detection, which has not been considered before.
- We design two novel modules, *i.e.*, the cross-layer feature aggregation module (CFA) and the cross-layer feature distribution module (CFD), which together allow efficient information communication across multiple layers.
- We develop the CFPN model based on the above two modules. It can bring consistent performance boost to a variety of backbones including VGG-16 [29], ResNet-18 [9] and ResNet-50 [9] for salient object detection. It establishes new state-of-the-arts on multiple benchmarks.

[1] In this paper, layers refer to the side-output features of the backbone.

## 2. Related Work

Early salient object detection methods usually rely on hand-crafted features and heuristic priors [4, 45, 30, 2], achieving only limited performance due to lack of high-level semantic information. Recently, benefiting from convolutional neural networks (CNNs), salient object detection enjoys much progress [11, 14, 47, 36].

Some deep saliency methods [14, 16, 31, 50] divide images into patches or superpixels, and extract single or multiple scales features from each patch or superpixel for determining whether the image regions are salient. Though better performance has been achieved than traditional methods, processing images in a patch-wise way ignores the essential spatial information of the whole image, which limits the accuracy for detecting the entire salient objects.

Some more effective models are developed based on fully convolutional networks (FCNs) [22]. Wang *et al*. [33] exploit low-level cues to generate guidance saliency maps by leveraging cascaded FCN. Liu *et al*. [20] develop a two-stage network which produces coarse saliency maps first and then integrates local context information to refine them recurrently and hierarchically. Hou *et al*. [11] introduce short connections into the HED [43] architecture, and predict salient objects based on aggregated saliency maps from each side-output. Wang *et al*. [34] propose to generate a coarse prediction map via FCN, and then refine it stage-wisely. Zhang *et al*. [47] utilize multi-level context information for accurate salient object detection with the HED network. In [35], Wang *et al*. propose to recurrently locate salient objects with local saliency cues. Zhang *et al*. [46] extract context-aware multi-level features and utilize a bi-directional gated structure to pass message between them.

Some works introduce the attention mechanism into the network design to exploit multi-level context information for saliency detection. For example, Zhang *et al*. [21] and Liu *et al*. [49] both devise attention guided networks in which multiple layer-wise attention is progressively integrated for saliency detection. Wang *et al*. [37] first extend regular attention mechanisms with multi-scale information to represent visual saliency contents, and then further improve salient object segmentation performance using salient edge information.

More recently, the feature pyramid networks (FPNs) [18] that are designed in a top-down manner have received growing attention in salient object detection. Liu *et al*. [19] propose a poolnet via plugging topmost level information into FPN fusion branch for detecting the salient objects jointly with the edge detection. Wu *et al*. [42] propose a cascaded partial decoder framework cascading high-level feature maps to refined the low-level features. We propose to detect salient objects by conducting cross-layer communication to enhance the progressive fusion of FPN branch for salient object detection.

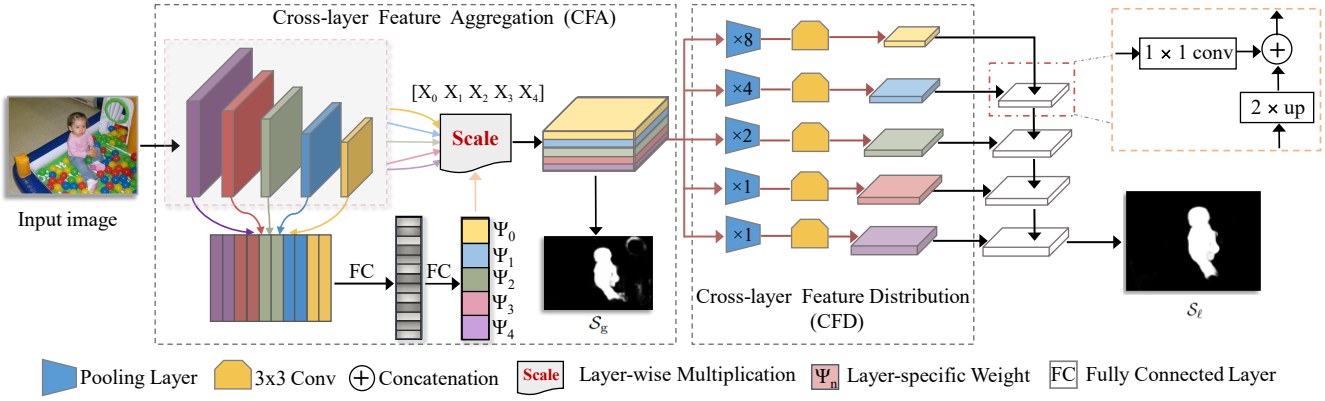



Figure 2: Overall framework of CFPN. It first extracts local sentations $(X_0, X_1, X_2, X_3, X_4)$ with backbone. Then, a cross-layer feature aggregation module (CFA) and a cross-layer feature distribution module (CFD) are inserted into the feature pyramid network (FPN) to explore the salient regions. Details of CFA are shown in Fig. 3 and Sec. 3.2; details of CFD are presented in Sec. 3.3.

# 3. Method

## 3.1. Overall Architecture

Fig. 2 shows the overall architecture of our Cross-layer Feature Pyramid Network (CFPN). It consists of two novel components, *i.e.* a Cross-layer Feature Aggregation module (CFA) and a Cross-layer Feature Distribution module (CFD). The CFA first adaptively generates a set of fusion weights for enhancing the original features at each layer by allowing information exchange among multiple layers. With this, the features are enhanced to have richer contexts. After CFA, the CFD allocates the aggregated features back to their corresponding layers via multi-scale pooling. Finally, facilitated by the distributed feature maps, CFPN gradually merges them in a top-down manner, similar to FPN, to produce the final saliency output.

## 3.2. Cross-layer Feature Aggregation

As described earlier, FPN based approaches often produce incomplete saliency maps due to gradual dilution of semantics during the progressive fusion. See Row 1 and 4 in Fig. 4 for illustration. Though recent works [19, 42] propose to aggregate the most top layer information into FPN fusion branch, this problem still exists and harms final results, as demonstrated in Column 2 and 5 in Fig. 4. In order to enable direct and more efficient communication among different layers, we propose to improve the fusion mechanism in FPN by aggregating all layer features simultaneously. Specifically, since the importance of different layer features largely depends on the image content, we devise a Cross-layer Feature Aggregation (CFA) module to adaptively predict a set of weights according to the importance of each level feature for aggregation. In this way, the features more useful for salient object detection will be promoted.

Denote the multi-level features output by the first pooling layer and the following four convolutional blocks of the ResNet [9] backbone as $\mathbf{X}_n, n \in \{0, 1, 2, 3, 4\}$. We first append a $1 \times 1$ convolutional layer at each level for dimension reduction, resulting in features with channel numbers $d_n \in \{64, 128, 256, 256, 256\}$. CFA then applies global average pooling at each level to squeeze its spatial information, and further concatenates channel-wise statistics from all the levels to integrate local and global contexts to construct multi-scale representations. Formally, given each level feature $\mathbf{X}_n \in \mathbb{R}^{H_n \times W_n \times d_n}$, CFA calculates the channel-wise global representation $\mathbf{Z} \in \mathbb{R}^{D \times 1}$ by

$$\mathbf{Z} = \Big\|_{n=1}^{N} \boldsymbol{z}_n = \Big\|_{n=1}^{N} \Big\{ \frac{1}{H_n \times W_n} \sum_{i=1}^{H_n} \sum_{j=1}^{W_n} \mathbf{X}_n(i,j) \Big\}, \tag{1}$$

where $\|$ is the concatenation function, $D = \sum_{n=1}^{N} d_n$ is the channel number of global representation $\mathbf{Z}$. $N$ refers to the overall index of local feature levels, and the pair-wise $(i, j)$ is the spatial coordinate of the feature map at each level.

We attempt to leverage the aggregated information $\mathbf{Z}$ to make each level features focus on salient regions instead of the overall feature maps. To this end, our CFA learns a layer-wise fusion weight $\mathbf{\Psi} \in \mathbb{R}^{1 \times N}$ by using a simple gating mechanism for $\mathbf{Z}$, i.e.,

$$\mathbf{\Psi} = \mathbf{W}_2\big(\mathrm{ReLU}(\mathbf{W}_1(\mathbf{Z}))\big). \tag{2}$$

Here $\mathbf{W}_1 \in \mathbb{R}^{D \times M}$ and $\mathbf{W}_2 \in \mathbb{R}^{M \times N}$ are two fully connected layers inspired by SENet [12], and $M$ denotes the transformed dimension of the global representation, which is set to 128 empirically. ReLU denotes the ReLU activation function. With the fusion weight $\mathbf{\Psi}$, we dynamically enhance each original layer feature by

$$\widetilde{\mathbf{X}}_n = \mathbf{X}_n * \Psi_n, \tag{3}$$

where $\Psi_n$ is the $n$-th element in the $\mathbf{\Psi}$, and $*$ means the scalar multiplication between $\mathbf{X}_n$ and $\Psi_n$. In this way, the

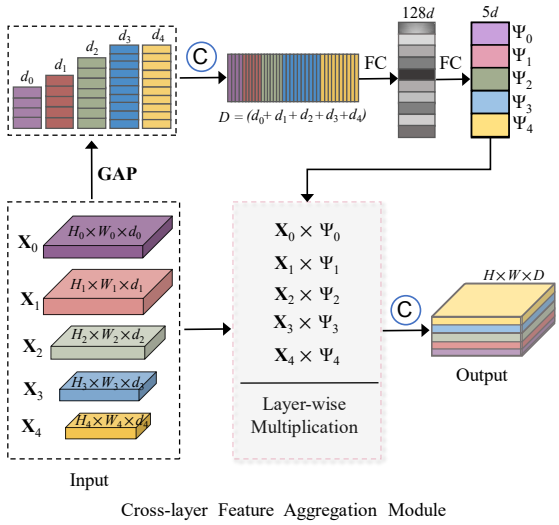



Figure 3: Detailed illustration of the proposed cross-layer feature aggregation module (CFA). $\Psi_n, n \in \{0, 1, 2, 3, 4\}$ is the learned layer-wise fusion weight for enhancing features per layer. GAP refers to global average pooling operation. ⓒ is the feature concatenation operation, and FC refers to the fully connected layer.

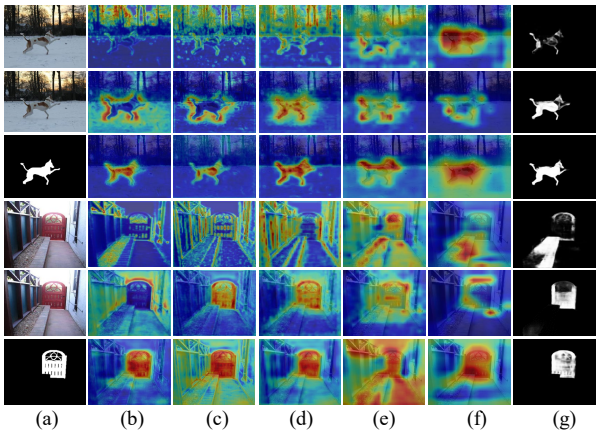



Figure 4: (a) Example input images and corresponding ground-truth labels. (b-f) Visualizations of progressive fusion feature maps at different levels from FPN (Row 1, 3), PoolNet [19] (Row 2, 4), and CFD (Row 3, 6). (g) Saliency maps generated from FPN (Row 1, 3), PoolNet [19] (Row 2, 4) and CFD (Row 3, 6), respectively. As can be seen, with our CFD, feature maps at each level contain richer contexts, which can more precisely highlight the whole salient objects (Row 3) and effectively suppress the over-predicted foreground regions (Row 6), compared to the villain FPN based decoder branch (Rows 1, 2, 4, 5).

adaptively enhanced multi-level features form a compact global image representation $\mathbf{F}$ for guiding accurate saliency detection. To be more specific, we first upsample $\widetilde{\mathbf{X}}_{2\sim4}$ to the same resolution as $\widetilde{\mathbf{X}}_0$ by bilinear interpolation, and then concatenate them to generate the global feature map $\mathbf{F}$. Formally, this process can be expressed as

$$\mathbf{F} = \widetilde{\mathbf{X}}_0 \oplus \widetilde{\mathbf{X}}_1 \oplus \mathbf{UP}(\widetilde{\mathbf{X}}_2) \oplus \mathbf{UP}(\widetilde{\mathbf{X}}_3) \oplus \mathbf{UP}(\widetilde{\mathbf{X}}_4), \quad (4)$$

where $\oplus$ refers to the concatenation operation, and $\mathbf{UP}$ denotes the upsampling function with bilinear interpolation.

### 3.3. Cross-layer Feature Distribution

Given the aggregated features from the previous CFA module, a direct method for producing the saliency map is to convolute the integrated feature with a new convolutional layer. Although this method can detect salient objects with richer contexts, the prediction is still not satisfactory by using such single stage inference, as shown in Fig. 6 (b). Instead, we propose to combine the aggregated feature maps with FPN, and infer salient regions in a stage-wise fusion manner. Unlike vanilla FPN, each layer feature now has access to the full spectrum of multi-level representation during the stage-wise fusion, thanks to the aggregation of multi-scale features by the CFA module. Thus, the aforementioned limitations of FPN are largely alleviated. To this end, we devise a Cross-layer Feature Distribution Module (CFD) to allocate multi-level features by performing multi-scale pooling over the aggregated feature $\mathbf{F}$. In this way, both semantics and salient details can be adaptively accessed at each level of fusion, which boosts the stage-wise fusion in FPN and helps better predict the whole salient objects, as shown in Rows 3 and 6 in Fig. 4.

Specifically, CFD first feeds $\mathbf{F}$ to the average pooling layers with pyramid downsampling rates to convert the aggregated features to different scale spaces. Taking the ResNet version of FPN as an example, the downsampling rates corresponding to levels $n \in \{0, 1, 2, 3, 4\}$ are $\{1, 1, 2, 4, 8\}$, respectively. Then, a $3 \times 3$ convolutional layer along with batch-normalization $(\mathrm{BN})$ and $\mathrm{ReLU}$ activation is appended after each downsampling operation to regenerate feature maps with channel numbers $\{64, 128, 256, 256, 256\}$ as $\widetilde{\mathbf{X}}_n, n \in \{0, 1, 2, 3, 4\}$, respectively. In this way, since the distributed feature maps at each fusion level simultaneously incorporate semantics and fine details, more discriminative and complementary representations can be well preserved along the progressive fusion path. The fusion effect is thus greatly enhanced for achieving more superior performance.

### 3.4. Model Training

Given the input image set $\mathcal{I}$ and its corresponding annotations $\mathcal{Y}$, we train our network with local and global saliency prediction jointly. This scheme can ensure salient objects uniformly highlighted and backgrounds suppressed, based on our comprehensive experiments.

With the CFA, we obtain the aggregated feature $\mathbf{F}$ with $(\frac{\mathrm{H}}{4}, \frac{\mathrm{W}}{4})$ size and $D$ channels. Then the global

saliency map $\mathcal{S}_{\mathrm{g}}$ is predicted with the readout function $\mathcal{R}_{\mathrm{g}}$: $\{\mathrm{Conv}(3\times 3,\ 128)\rightarrow \mathrm{BN}\rightarrow \mathrm{ReLU}\rightarrow \mathrm{Conv}(1\times 1,\ 1)\rightarrow \mathrm{upsampling}(\mathrm{H},\mathrm{W})\rightarrow \mathrm{sigmoid}\}$. For predicting the local saliency map $\mathcal{S}_{\ell}$, after learning the local representation $\mathbf{L}$ from CFD, the prediction function $\mathcal{R}_{\ell}$: $\{\mathrm{Conv}(1\times 1,\ 1)\rightarrow\ \mathrm{upsampling}(\mathrm{H},\mathrm{W})\rightarrow \mathrm{sigmoid}\}$, is used to produce $\mathcal{S}_{\ell}$ directly. According to $\{\mathcal{S}_{\ell},\mathcal{S}_{\mathrm{g}}\}$, our network is trained by formulating the loss function

$$\mho = \mathcal{L}_{bce}(\mathcal{S}_{\mathrm{g}},\mathcal{Y}|\Theta_{\mathrm{g}}) + \mathcal{L}_{bce}(\mathcal{S}_{\ell},\mathcal{Y}|\Theta_{\ell}), \tag{5}$$

where the network parameter $\Theta=\{\Theta_{\ell},\Theta_{\mathrm{g}}\}$ is used to generate the saliency maps $\{\mathcal{S}_{\ell},\mathcal{S}_{\mathrm{g}}\}$. The $\mathcal{L}_{bce}$ is the balanced binary cross entropy loss

$$\begin{aligned}\mathcal{L}_{bce}(\Theta) = &-\beta \sum_{j\in\mathcal{Y}_{+}} \log \Pr(\mathcal{Y}^{j}=1|\Theta) \\ &-(1-\beta)\sum_{j\in\mathcal{Y}_{-}} \log \Pr(\mathcal{Y}^{j}=0|\Theta),\end{aligned} \tag{6}$$

where $j$ denotes pixel coordinate, and $\mathcal{Y}_{+}$, $\mathcal{Y}_{-}$ are the foreground and background label sets, respectively. $\beta$ is the loss weight which is defined as $\beta=\mathcal{Y}_{+}/\mathcal{Y}_{-}$. The salient confidence score $\Pr=(1+e^{-\mathcal{S}})^{-1}$.

# 4. Experiments

## 4.1. Settings

**Datasets** To evaluate the proposed approach, we experiment on six saliency detection benchmark datasets, including ECSSD [44], PASCAL-S [17], DUT-OMRON [45], HKU-IS [15], SOD [25] and DUTS-test [32], which respectively contain 1,000, 850, 5,168, 4,447, 300 and 5,019 natural complex images with manually labeled pixel-wise ground-truths.

**Implementation Details** We perform all experiments using the adam [13] optimizer with initial learning rate 5e-5, 0.9 momentum, 5e-4 weight decay, and batch size 14. Following previous works [19, 42, 21, 49, 46, 34], we use the training set of DUTS [32] dataset to train the proposed model. The training samples are augmented through random rotation and horizontal flipping. The backbone (VGG-16 [29], ResNet-18 [9], and ResNet-50 [9]) parameters of our network are initialized with the corresponding models pretrained on ImageNet [5] and the rest are randomly initialized. In both training and testing phrases, input images are resized to $384\times 384$. Different from some recent saliency models trained with extra supervision constraints (e.g., boundary [27, 35], edge [7, 37, 19]) or post processing operations (e.g., CRF [11, 21]), our network simply uses pixel-level saliency annotations, with no extra processes used when generating final saliency maps.

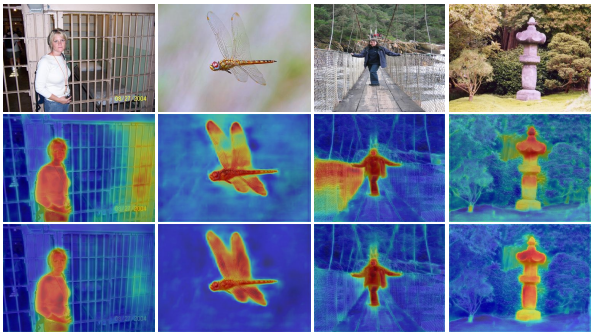


Figure 5: Visualizing feature maps generated by directly aggregating the original multiple layer features (Row 2) and the CFA module (Row 3). Obviously, feature maps from CFA can more precisely capture the positions and contours of salient objects (Row 3).

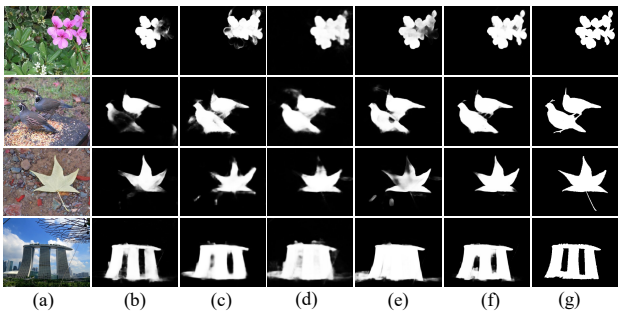
**Evaluation Metrics** We adopt three metrics: precision-recall (PR) curves, F-measure, and mean absolute error (MAE) as our evaluation metrics. For F-measure, we report the maximum $F_{\beta}$ (MaxF) for evaluating our method and state-of-the-art approaches, as similar to recent studies [47, 49, 11, 35, 48, 23, 6, 19].

## 4.2. Ablation Studies

We first analyze the contributions of each module in our method, namely CFA and CFD, to overall performance. Then, different configurations of feature enhancement strategies are compared to validate our CFA design. At last, by allocating different numbers of layer features over the aggregartion feature map, we verify the effect of CFD design on improving progressive fusion for detecting salient regions. All ablation experiments are conducted with ResNet-50 backbone on DUT-OMRON [45], PASCAL-S [17] and DUTS-TE [32] datasets.

**Effectiveness of CFA and CFD** We compare three variants of backbone with the FPN baseline: *w/* CFA, *w/* CFD. Fig. 4, Fig. 5, Fig. 6 show some visualized results, and Tab. 1 shows MaxF and MAE scores of CFA and CFD on three challenging datasets.

- ***w/* CFA**: By comparing results of backbone Res50 (Row 1 in Tab. 1, ***w/o* CFA**), the addition of CFA (Row 2 in Tab. 1) obviously brings performance gain in terms of both MaxF and MAE scores. Besides, compared to Row 2 in Tab. 1, CFA consistently outperforms the vanilla FPN, with a margin of 2.1% and 2.9% on DUTS-O, PASCAL-S dataset w.r.t. MaxF, respectively. This validates the effectiveness of our dynamic cross-layer feature aggregation strategy.
  From visualization results in Fig. 5, when comparing Row 2 (w/o CFA) and Row 3 (w CFA), feature maps after CFA provide more discriminative information for



(a) (b) (c) (d) (e) (f) (g)

Figure 6: Visualization of saliency maps predicted by aggregated feature **F**, FPN based models, and our method. (a) Source images. (b-f) Results of backbone + CFA, CASNet [42], PiCANet [21], PoolNet [19], backbone + CFA + CFD. (g) Ground Truth.

| No. | Module | DUT-O[45] | | PASCAL-S[17] | |
|---|---|---|---|---|---|
| | | MaxF ↑ | MAE ↓ | MaxF ↑ | MAE ↓ |
| 1 | Res50 | 0.761 | 0.084 | 0.833 | 0.128 |
| 2 | Res50 + FPN | 0.796 | 0.065 | 0.845 | 0.087 |
| 3 | Res50 + CFA | **0.817** | **0.061** | **0.874** | **0.079** |
| 4 | Res50 + CFA + CFD | **0.834** | **0.053** | **0.886** | **0.072** |

Table 1: Ablation analysis w.r.t. effectiveness of CFA/CFD. Res50 is the ResNet-50 backbone. CFA and CFD in our method are important for improving performance. Best and second best results are shown in **black** and **red**, respectively.

distinguishing foregrounds from clutter backgrounds, and thus can better locate the entire salient object than those without CFA. Moreover, by adaptively aggregating multi-layer features, the CFA greatly improves the quality of generated global saliency maps, as shown in Fig. 6 (b). These results clearly demonstrate that saliency detection benefits from dynamic feature aggregation over information exchanging across multiple layer features.

- ***w/* CFD**: Comparing Row 3 and 4 in Tab. 1, collaborating with CFD (Row 4), the MaxF scores are improved with a margin of 1.7% , 1.2% on DUT-O and PASCAL-S datasets, and the MAE values are decreased from 0.061 to 0.053 for DUT-O dataset, from 0.079 to 0.072 for PASCAL-S dataset, respectively. Moreover, by comparing results of FPN, applying both CFA and CFD greatly improve performance in both MaxF and MAE values.

  Fig. 4 gives the visualization feature maps at each level after CFD. Obviously, by comparing Rows 3 and 6 (w CFD) with Rows 1, 2, 4, and 5 (w/o CFD), the distributed feature maps at each fusion level provide rich semantics and clear object boundaries, ensuring that entire salient objects can be segmented with sharp object boundaries (Row 3 and 6 (g)).

  Fig. 6 (f) and (b) gives Some corresponding saliency maps between w/ and w/o CFD. Clearly, inaccurate saliency results, e.g. over-predicted and incomplete objects, blurred object boundaries, get greatly improved by collaborating with the CFD. These results consistently demonstrate the effectiveness of CFD.

| Module | DUT-O[45] | | PASCAL-S[17] | | DUTS-TE[32] | |
|---|---|---|---|---|---|---|
| | MaxF ↑ | MAE ↓ | MaxF ↑ | MAE ↓ | MaxF ↑ | MAE ↓ |
| **(A)** | 0.803 | 0.069 | 0.861 | 0.081 | 0.863 | 0.049 |
| **(B)** | 0.811 | 0.064 | 0.868 | 0.078 | 0.870 | 0.047 |
| **(C)** | 0.813 | 0.062 | 0.870 | 0.079 | 0.872 | 0.047 |
| **(D)** | **0.817** | **0.061** | **0.874** | **0.079** | **0.875** | **0.045** |

Table 2: Ablation analysis w.r.t. different configurations of CFA. Design of CFA achieves better performance than other settings. Best results are shown in **red**.

**Configurations of CFA** We here analyze the effectiveness of our CFA design, which simultaneously considers multi-level features for adaptive layer-wise reweighting during aggregation. We compare our approach against the following baselines, including:

(A) **No reweighting:** The feature maps from each layer are directly concatenated, followed by a $1 \times 1$ conv layer for saliency map prediction.

(B) **Non-learnable reweighting:** We use global average pooling (GAP) on each level features to obtain the layer-wise weights and multiply them with the original features for aggregation before producing $\mathcal{S}_g$.

(C) **Independent layer-wise reweighting:** Similar to (B), we apply GAP on each level features, followed by two fully connected layers before multiplying with the original features. This is performed **independently** on each level before concatenation.

(D) **Collaborative layer-wise reweighting:** We apply our CFA module to learn a set of layer-wise weights by simultaneously considering all the information among different layers for aggregation, as discussed in 3.2.

Tab. 2 reports the qualitative results of the above settings. As can be observed, both **(B)** and **(C)** significantly outperform the method **(A)**. This confirms that dynamically leveraging multi-level features is crucial for saliency detection. However, **(B)** and **(C)** give inferior performance to **(D)**, because the two designs reweight each level features by viewing global weights from themselves independently, which ignores the channel interdependencies among different levels. On the contrary, with collaborative layer-wise reweighting, CFA obviously achieves better performance for predicting $\mathcal{S}_g$. These results indicate that the design of CFA plays an important role in boosting saliency performance.

**Configurations of CFD** To be better illustrating the distribution process in CFD, we allocate the aggregated feature **F** to different numbers of level features. Tab. 3 reports

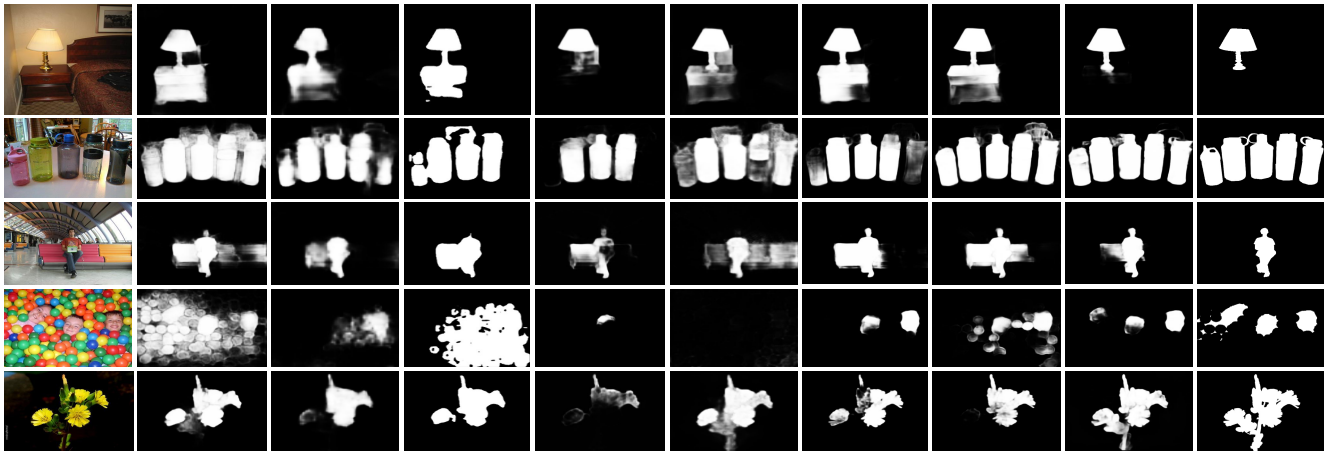



Figure 7: Comparison of saliency maps generated by our method and previous state-of-the-arts. It can be seen that our method can not only locate the entire foreground salient objects but also effectively suppress cluster backgrounds, even for some challenging scenes. Best viewed in color.

| No. | Settings | PASCAL-S[17] | | DUTS-TE[32] | |
|---|---|---|---|---|---|
| | | MaxF ↑ | MAE ↓ | MaxF ↑ | MAE ↓ |
| 1 | **(D)** | 0.874 | 0.079 | 0.875 | 0.045 |
| 2 | $\{\widetilde{\mathbf{X}}_0\}$ | 0.880 | 0.075 | 0.882 | 0.040 |
| 3 | $\{\widetilde{\mathbf{X}}_0, \widetilde{\mathbf{X}}_1\}$ | 0.879 | 0.076 | 0.885 | 0.040 |
| 4 | $\{\widetilde{\mathbf{X}}_0, \widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2\}$ | 0.881 | 0.074 | 0.887 | 0.038 |
| 5 | $\{\widetilde{\mathbf{X}}_0, \widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2, \widetilde{\mathbf{X}}_3\}$ | 0.883 | 0.073 | 0.889 | 0.038 |
| 6 | $\{\widetilde{\mathbf{X}}_0, \widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2, \widetilde{\mathbf{X}}_3, \widetilde{\mathbf{X}}_4\}$ | **0.886** | **0.072** | **0.896** | **0.035** |

Table 3: Ablation analysis of CFD with different distribution configurations. **(D)** refers to w/o CFD module defined in Tab. 2. Each level feature in CFD contributes a lot to the progressive fusion. Best results are highlighted in **red**.

the corresponding comparison results in terms of MaxF and MAE values on two challenging datasets. By comparing results of Row 1 in Tab. 3, the CFD module (Rows 2∼6) contributes a lot to produce better saliency results. This further demonstrates that the stage-wise fusion performs better than the single stage fusion for saliency detection. Besides, by distributing $\mathbf{F}$ into 1∼5 levels for progressive fusion respectively, the performance is gradually improved, illustrating that each level feature in CFD plays an important role for the progressive fusion.

### 4.3. Comparison with State-of-the-Arts

We compare our proposed method with 14 deep saliency detection methods, including DCL [16], DSS [11], NLDF [23], Amulet [47], SRM [34], DGRL [35], R3Net [6], BMPM [46], PAGR [49], PiCANet [21], AFNet [7], BASNet [27], CASNet [42], and PoolNet [19]. For fair comparison, we cite the public comparison results provided by [24], which generate saliency maps from the source code released by the authors or directly provided by them. We evaluate all the competitors with the same evaluation code.

**Visual Comparison** Fig. 7 shows visual comparisons of the proposed model (Ours) with previous state-of-the-art methods. We can clearly see that our model highlights salient objects closest to the ground-truth maps in various challenging scenarios, including images with cluster backgrounds and foregrounds (Row 3, 4), object having similar appearance to background (Row 1, 3, 4), multiple instances of the same object (Row 2, 5), and objects occluded by background objects (Row 3, 4). More importantly, our model can well segment the entire objects (Row 1, 2, 3, 5) with clear salient object boundaries (Row 1, 2, 3, 4, 5), demonstrating the effectiveness of the proposed CFPN.

**F-measure and MAE Comparison** Tab. 4 reports the MaxF and MAE scores of our method using different backbones (VGG-16 [29], ResNet-18 [9], and ResNet-50 [9]) compared with other methods. Obviously, CFPN achieves excellent results on all the datasets with the similar backbones across the metrics. In particular, with both VGG-16 [29] and ResNet-50 [9] backbones, CFPN shows significantly improved $F_\beta$-max scores compared with the second best PoolNet [19], on the more challenging benchmarks PASCAL-S (**VGG-16**: 0.874 *vs* 0.857; **ResNet-50**: 0.886 *vs* 0.863), DUTS-TE (**VGG-16**: 0.885 *vs* 0.876; **ResNet-50**: 0.896 *vs* 0.886), and HKUIS (**VGG-16**: 0.937 *vs* 0.928; **ResNet-50**: 0.940 *vs* 0.934). More importantly, when using ResNet-18 [9] as backbone, our CFPN not only outperforms all the previous VGG backbone approaches significantly, but also beats most of the ResNet-50 based methods, especially on the more challenging datasets including PASCAL-S, SOD, and DUTS-TE. These results clearly illustrate the superior performance and robustness of CFPN.

| Methods | Backbone | ECSSD [44] | | PASCAL-S [17] | | DUTS-TE [32] | | HKU-IS [15] | | SOD [25] | | DUT-OMRON [45] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MaxF ↑ | MAE ↓ | MaxF ↑ | MAE ↓ | MaxF ↑ | MAE ↓ | MaxF ↑ | MAE ↓ | MaxF ↑ | MAE ↓ | MaxF ↑ | MAE ↓ |
| **VGG backbone** | | | | | | | | | | | | | |
| DCL CVPR2016 [16] | VGG-16 | 0.890 | 0.088 | 0.805 | 0.125 | 0.782 | 0.088 | 0.885 | 0.072 | 0.823 | 0.141 | 0.739 | 0.097 |
| DSS CVPR2016 [11] | VGG-16 | 0.916 | 0.053 | 0.836 | 0.096 | 0.825 | 0.057 | 0.911 | 0.041 | 0.844 | 0.121 | 0.771 | 0.066 |
| NLDF CVPR2017 [23] | VGG-16 | 0.905 | 0.063 | 0.831 | 0.099 | 0.812 | 0.066 | 0.902 | 0.048 | 0.841 | 0.124 | 0.753 | 0.080 |
| Amulet ICCV2017 [47] | VGG-16 | 0.915 | 0.059 | 0.837 | 0.098 | 0.778 | 0.085 | 0.895 | 0.052 | 0.806 | 0.141 | 0.742 | 0.098 |
| BMPM CVPR2018 [46] | VGG-16 | 0.929 | 0.045 | 0.862 | 0.074 | 0.851 | 0.049 | 0.921 | 0.039 | 0.855 | 0.107 | 0.774 | 0.064 |
| PAGR CVPR2018 [49] | VGG-19 | 0.927 | 0.061 | 0.856 | 0.093 | 0.855 | 0.056 | 0.918 | 0.048 | - | - | 0.771 | 0.071 |
| PiCANet CVPR2018 [21] | VGG-16 | 0.931 | 0.047 | 0.868 | 0.077 | 0.851 | 0.054 | 0.921 | 0.042 | 0.853 | 0.102 | 0.794 | 0.068 |
| AFNet CVPR2019 [7] | VGG-16 | 0.935 | 0.042 | 0.868 | 0.071 | 0.862 | 0.046 | 0.923 | 0.036 | 0.856 | 0.109 | 0.797 | 0.057 |
| PoolNet CVPR2019 [19] | VGG-16 | 0.936 | 0.047 | 0.857 | 0.078 | 0.876 | 0.043 | 0.928 | 0.035 | 0.859 | 0.115 | 0.817 | 0.058 |
| **Ours (VGG)** | VGG-16 | **0.943** | **0.040** | **0.874** | **0.071** | **0.885** | **0.038** | **0.937** | **0.031** | **0.870** | **0.097** | **0.829** | **0.054** |
| **ResNet backbone** | | | | | | | | | | | | | |
| SRM ICCV2017 [34] | ResNet-50 | 0.917 | 0.054 | 0.847 | 0.085 | 0.827 | 0.059 | 0.906 | 0.046 | 0.843 | 0.127 | 0.769 | 0.069 |
| DGRL CVPR2018 [35] | ResNet-50 | 0.922 | 0.041 | 0.854 | 0.078 | 0.829 | 0.056 | 0.910 | 0.036 | 0.845 | 0.104 | 0.774 | 0.062 |
| R3Net IJCAI2018 [6] | ResNeXt | 0.931 | 0.046 | 0.845 | 0.097 | 0.828 | 0.059 | 0.917 | 0.038 | 0.836 | 0.136 | 0.792 | 0.061 |
| PiCANet CVPR2018 [21] | ResNet-50 | 0.935 | 0.047 | 0.881 | 0.087 | 0.860 | 0.051 | 0.919 | 0.043 | 0.858 | 0.109 | 0.803 | 0.065 |
| BASNet CVPR2019 [19] | ResNet-34 | 0.942 | 0.037 | 0.854 | 0.076 | 0.860 | 0.047 | 0.928 | 0.032 | 0.851 | 0.114 | 0.805 | 0.056 |
| CASNet CVPR2019 [42] | ResNet-50 | 0.939 | 0.037 | 0.864 | 0.072 | 0.865 | 0.043 | 0.925 | 0.034 | - | - | 0.797 | 0.056 |
| PoolNet CVPR2019 [19] | ResNet-50 | 0.940 | 0.042 | 0.863 | 0.075 | 0.886 | 0.040 | 0.934 | 0.032 | 0.867 | 0.100 | 0.830 | 0.055 |
| **Ours (Res18)** | ResNet-18 | **0.942** | **0.039** | **0.879** | **0.074** | **0.887** | **0.039** | **0.933** | **0.032** | **0.872** | **0.085** | **0.821** | **0.055** |
| **Ours (Res50)** | ResNet-50 | **0.948** | **0.035** | **0.886** | **0.072** | **0.896** | **0.035** | **0.940** | **0.029** | **0.873** | **0.083** | **0.834** | **0.053** |

Table 4: Comparisons of max F-measure and MAE values on VGG [29] and ResNet [9] backbones are reported. Results of our method are shown in **blue**, **black**, and **red**, respectively. With different backbones, the proposed method consistently achieves better performance than the previous state-of-the-arts. Best viewed in color.

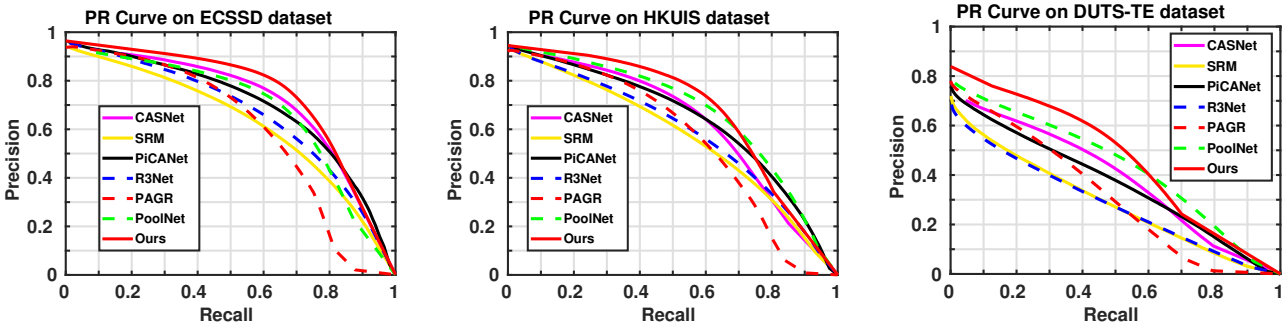



Figure 8: Precision and recall curves on ECSSD [44], HKUIS [14], and DUTS-TE [32] datasets. The proposed method outperforms previous state-of-the-arts on all the datasets. Best viewed in color.

**PR Curves Comparison** We also give the precision-recall curves in Fig. 8. Due to limited space, we simply show the PR curves of the previous methods implemented with ResNet-50 backbone over three widely used datasets. As can be seen, the PR curves of our CFPN, represented by the straight red lines, consistently outperform all other previous models over all datasets. These results convincingly demonstrate the effectiveness of our method.

## 5. Conclusion

In this paper, we identify the limitation of FPN based saleincy methods (i.e., *indirect* information propagation between deeper and shallower layers) and presented a novel architecture, CFPN, for salient object detection. It consists of two essential modules: a cross-layer feature aggregation module and a cross-layer feature distribution module. Benefiting from these two collaborative modules, efficient information communication across multiple layers is conducted, which reduces the information loss during FPN stage-wise fusion, and thus leads to more accurate saliency results. Comprehensive experiments on popular saliency detection benchmarks demonstrate the effectiveness and robustness of the proposed CFPN.

## References


[1] Md Amirul Islam, Mahmoud Kalash, and Neil DB Bruce. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *CVPR*, pages 7142–7150, 2018.

[2] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational Visual Media*, pages 1–34, 2014.

[3] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *ECCV*, pages 236–252, 2018.

[4] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *TPAMI*, 37(3):569–582, 2015.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[6] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng. $R^3$Net: Recurrent residual refinement network for saliency detection. In *IJCAI*, pages 684–690, 2018.

[7] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *CVPR*, pages 1623–1632, 2019.

[8] Wei Feng, Ruize Han, Qing Guo, Jianke Zhu, and Song Wang. Dynamic saliency-aware regularization for correlation filter-based object tracking. *IEEE Transactions on Image Processing*, 28(7):3232–3245, 2019.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2015.

[10] Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han. Online tracking by learning discriminative saliency map with convolutional neural network. In *ICML*, pages 597–606, 2015.

[11] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. In *CVPR*, pages 5300–5309, 2017.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015.

[15] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015.

[16] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *CVPR*, pages 478–487, 2016.

[17] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, pages 280–287, 2014.

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

[19] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, 2019.

[20] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, pages 678–686, 2016.

[21] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, pages 3089–3098, 2018.

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *TPAMI*, 39(4):640–651, 2014.

[23] Zhiming Luo, Akshaya Kumar Mishra, Andrew Achkar, Justin A Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *CVPR*, pages 6593–6601, 2017.

[24] Mengyang Feng. Evaluation toolbox for salient object detection. `https://github.com/ArcherFMY/sal_eval_toolbox`, 2018.

[25] Vida Movahedi and James H Elder. Design and perceptual validation of performance measures for salient object segmentation. In *CVPRW*, pages 49–56, 2010.

[26] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499, 2016.

[27] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489, 2019.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[29] K. Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2018.

[30] Jingdong Wang, Huaizu Jiang, Zejian Yuan, Ming Ming Cheng, Xiaowei Hu, and Nanning Zheng. Salient object detection: A discriminative regional feature integration approach. *IJCV*, 123(2):1–18, 2017.

[31] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *CVPR*, pages 3183–3192, 2015.

[32] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017.

[33] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *ECCV*, pages 825–841, 2016.

[34] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *ICCV*, pages 4019–4028, 2017.

[35] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *CVPR*, pages 3127–3135, 2018.

[36] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, and Haibin Ling. Salient object detection in the deep learning era: An in-depth survey. *arXiv preprint arXiv:1904.09146*, 2019.

[37] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji. Salient object detection with pyramid attention and salient edges. In *CVPR*, pages 1448–1457, 2019.

[38] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *TPAMI*, 39(11):2314–2320, 2015.

[39] Jianbing Shen Wenguan Wang and Haibin Ling. A deep network solution for attention and aesthetics aware photo cropping. *IEEE TPAMI*, abs/1612.03144, 2018.

[40] Jianbing Shen Wenguan Wang and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *CVPR*, pages 3395–3402, 2015.

[41] Runmin Wu, Mengyang Feng, Wenlong Guan, Dong Wang, Huchuan Lu, and Errui Ding. A mutual learning method for salient object detection with intertwined multi-supervision. In *CVPR*, pages 8150–8159, 2019.

[42] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, pages 3907–3916, 2019.

[43] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, pages 1395–1403, 2015.

[44] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162, 2013.

[45] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013.

[46] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *CVPR*, pages 1741–1750, 2018.

[47] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, pages 202–211, 2017.

[48] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, pages 212–221, 2017.

[49] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *CVPR*, pages 714–722, 2018.

[50] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *CVPR*, pages 1265–1274, 2015.

[51] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *CVPR*, pages 3085–3094, 2019.