Holistic and Deep Feature Pyramids for Saliency Detection

Shizhong Dong [1,2]
sz.dong@siat.ac.cn
Zhifan Gao[3]
gaozhifan@gmail.com
Shanhui Sun[4]
shanhuis@curacloudcorp.com
Xin Wang[4]
xinw@curacloudcorp.com
Ming Li [1,2]
ming.li1@siat.ac.cn
Heye Zhang[1]
hy.zhang@siat.ac.cn
Guang Yang[5]
g.yang@imperial.ac.uk
Huafeng Liu[6]
liuhf@zju.edu.cn
Shuo Li[3]
slishuo@gmail.com

[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
Shenzhen, China

[2] Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences
Shenzhen, China

[3] Western University
London, Canada

[4] Curacloud Corporation
Seattle, USA

[5] Imperial College London
London, UK

[6] Zhejiang University
Hangzhou, China

Abstract

Saliency detection has been increasingly gaining research interest in recent years since many computer vision applications need to derive object attentions from images in the first steps. Multi-scale awareness of the saliency detector becomes essential to find thin and small attention regions as well as keeping high-level semantics. In this paper, we propose a novel holistic and deep feature pyramid neural network architecture that can leverage multi-scale semantics in feature encoding stage and saliency region prediction (decoding) stage. In the encoding stage, we exploit multi-scale and pyramidal hierarchy of feature maps via the densely connected network with variable-size dilated convolutions as well as a pyramid pooling. In the decoding stage, we fuse multi-level feature maps via up-sampling and convolution. In addition, we utilize the multi-level deep supervision via plugging in loss functions at every feature fusion level. Multi-loss supervision regularizes weights searching space among different tasks minimizing overfitting and enhances gradient signal during backpropagation, and thus enables us training the network from scratch. This architecture builds an inherent multi-level semantic pyramidal feature maps at different scales and enhances model's capability in the saliency detection task. We validated our approach on six benchmark datasets and compared with

Corresponding authors: Zhifan Gao (gaozhifan@gmail.com) and Heye Zhang (hy.zhang@siat.ac.cn)
The National Natural Science Foundation of China (No: 61525106, 61427807,61771464), shenzhen innovation funding (JCYJ20170307165309009, JCYJ20170413114916687,SGLH20161212104605195)

eleven state-of-the-art methods. The results demonstrated that the design effectiveness and our approach outperformed the compared methods.

1 Introduction

Salient region detection in visual scenes aims to seek attention of human visual system, which is essential in cognitive psychology and neurobiology that how humans perceive and process the stimuli from sights [19, 27]. Visual Saliency is a fundamental step for many computer vision tasks, such as image understanding and cognition, explainable computer vision (like static or dynamic scene captioning), and visual question and answer [3, 4, 8, 20, 33]. Recently, the saliency detection approaches based on convolutional neural network (CNN) outperformed those based on hand-engineered features [5, 23]. In the development of CNN-based approaches, the bottom-up pathway (feedforward encoding computation) computes a feature hierarchy consisting of feature maps at multiple scales using either max or average down-sample pooling. The top-down pathway (prediction decoding computation) predicts salient regions consisting of convolution and up-sampling operations hierarchically from feature maps generated by the bottom-up pathway. Predictions are made at the last level or made independently at different levels. Independent predictions are not aware of multi-scale semantics. The drawback of only using the high-level feature maps is not able to extract thin or small salient regions [6], as well as lead to heavily blurred region boundaries. Only considering the low-level feature maps leads to the algorithm losing larger contextual semantics. To bring in multi-scale awareness in the network, one popular approach is to jointly make prediction across different level feature maps. However, high-level feature maps do not have low-level information and vice versa. In contrast, the feature pyramid considers a large range of scale changes in feature space at the same time and thus the saliency detector is strongly invariant to object's scale changes.

In this work, we distill above insights of the feature pyramid and propose a novel holistic feature pyramid architecture implemented in both bottom-up and top-down pathways. We coin our network as a holistic and deep feature pyramid network for saliency detection. Figure 1 illustrates the overview of our approach. In the bottom-up pathways, we propose a multi-level and pyramidal hierarchical convolution architecture for feature abstraction for appropriately utilizing the low-level and high-level semantic information. In each level, we use a densely connected networks with dilated convolution to extract the information at the corresponding semantic degree. Varied dilated rates in different levels can provide a pyramidal hierarchy to increase the scale of feature extraction with enlarging the receptive fields. This can help to search the salient components within feature maps in the multi-scale space. Besides, we apply the pyramid pooling module (PPM) to further represent the highest-level feature maps, for extracting the global contextual information in the largest receptive field. Then we fuse multi-level feature maps via up-sampling and concatenation in the top-down pathway to assure that the predicted saliency map in every level contains the semantic information from all higher-level features. This leads to our final saliency prediction being aware of high-level semantic information and low-level fine-grained information.

Densely connected convolution blocks are able to forward the signal in the forward pass and the gradient signal in the backward pass during training (as discussed in [7]). However, the other part of network experiences gradient vanishing problem due to deep layer structure. To enhance gradient signal in the backpropagation procedure, we utilize multi-level deep supervision via plugging in loss functions at every feature fusion level. Another advantage

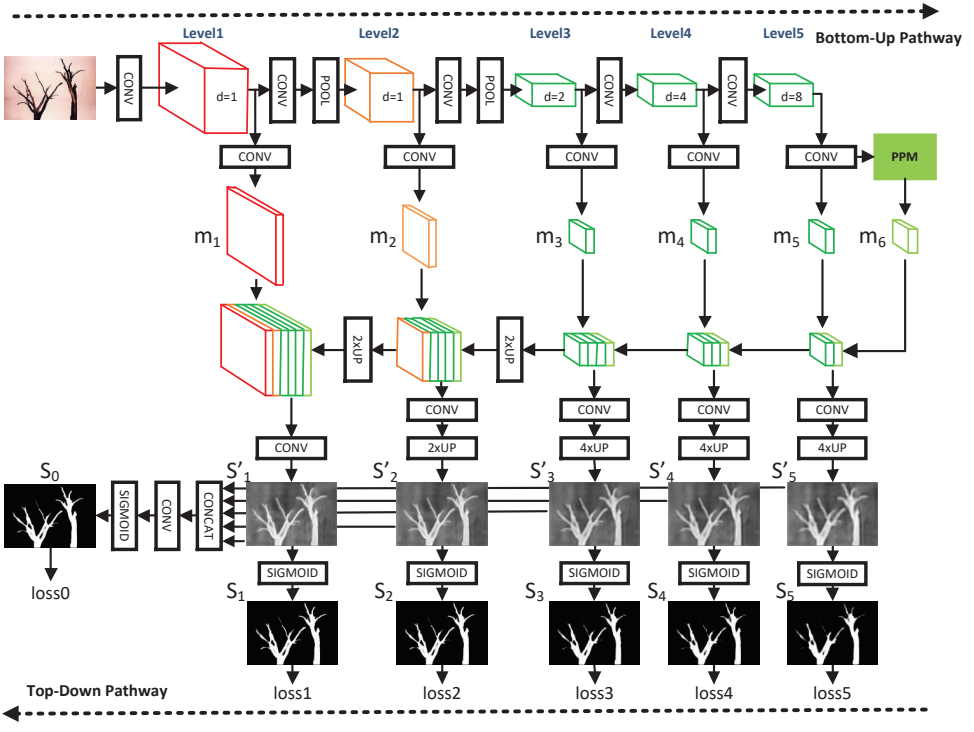

Figure 1: The architecture of our holistic and deep feature pyramid neural network for saliency detection. Our network contains five blocks with dilated densely-connected convolutional network (referred as to dilated dense block, DDB) and a pyramid pooling module (PPM [32]). We place a convolutional layer in the middle of each adjacent DDBs to compress the channels of DDB's output. The compress rate is 0.5. We just place a 2×2 pooling layer behind the first two DDBs. Each DDB includes 12 dilated convolution layers. Each dilated convolution layer is a combination of 'BN+ReLu+Dilated Conv'. The output of each dilated convolution layer has same 12 channels (termed as growth rate in DenseNet [7]). "d" denotes the dilated rate. The dilated rate of dilated convolution in five blocks are 1,1,2,4 and 8 separately. $m_1 \sim m_5$ are the feature maps extracted from the outputs of five DDBs. Each has 16 channels. m_6 is the output of PPM. It has 4 channels. $m_1 \sim m_6$ are hierarchically concacted from m_6 to m_1. These five concatenation feature maps are adopted to generate five final feature maps $S_1^{'} \sim S_5^{'}$. All the final feature maps are upsample to the same resolution as input image 256×256. The upsample method we used is bilinear interpolation. After that, these five final feature maps are feed to sigmoid to produce five saliency maps $S_1 \sim S_5$. The final saliency map S_0 is aggregated from five final feature maps.

of deep supervision is that multi-tasks (multi-level prediction) compete each other and thus regularize each other to minimize over-fitting problem. The deep supervision facilitates to achieve a good decision hypothesis (good generalization), and enables us to train the network from scratch.

The contributions of our approach is summarized as follows:

1. We developed a deep bi-directional pyramid neural network to build a holistic representation of different scale salient features. Because the size or shape of salient region might change sharply over different images, it is difficult to capture different salient

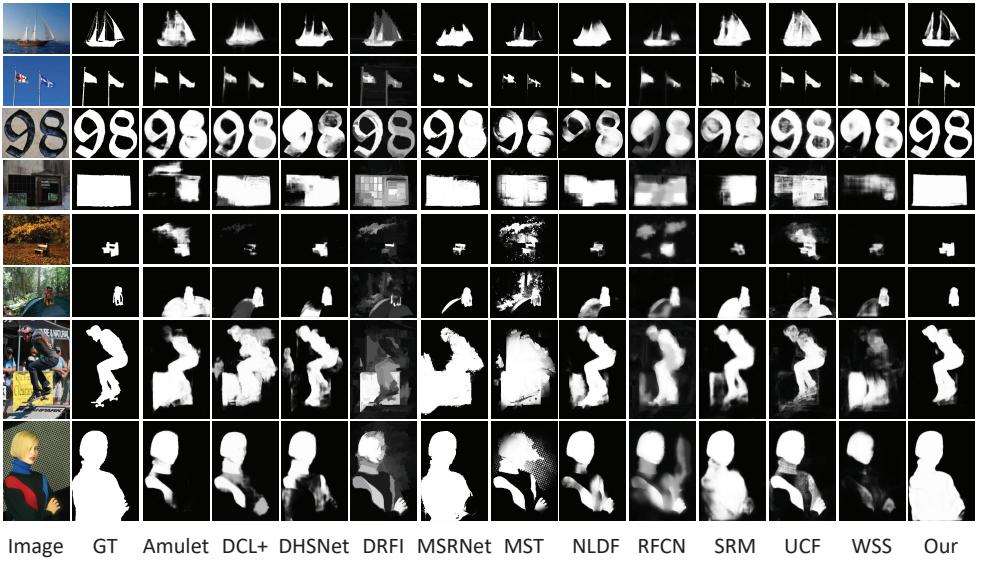

Figure 2: Examples of saliency detection results produced by our approach and the compared state-of-the-art methods. "GT" represents ground truth. Note that our approach is able to draw attention to thin and small structure (e.g. the flag pole). The results show that our approach is superior to the state-of-the-art method.

regions using few receptive field. In our deep pyramid neural network, we use five levels of holistic features that have different receptive fields to detect multi-scale salient regions from images. Meanwhile, the multi-level deep pyramid hierarchical architecture in our network can effectively extract and fuse the saliency information from every abstraction level, i.e. high-level features can locate salient region and low-level features can keep more spatial detailed information. Concatenating different levels feature will produce more precisely saliency map (see Figure 1).

2. The well-designed architecture of our network is able to perceive the accurate locations of salient objects, without the need to extract image features in the very high-level scale. It can overcome the side effect commonly existing in the previous studies. The side effect comes from that previously very deep networks always applied to extract the contextual information to localize salient objects. The very deep networks will largely decrease the resolution of feature maps at high level (commonly due to pooling), and further lead to missing of detailed information within these high-level features. The missing of detailed information will also introduce noises in the subsequent upsampling of the high-level feature maps. Thus, we design a holistic and pyramid network with multiple dense blocks with dilated convolution for simultaneously preserving the detailed information and capturing sufficient semantic information.

3. We propose an effective training scheme (i.e. deep supervision using multiple losses in every level) in our deep pyramid architecture, which can overcome the gradient vanishing and over-fitting problem by construct the direct feedback to every convolution layer from the ground truth. Every dilated convolution layer applied in our network can directly feed the information to the output of the corresponding dense block, which

further feed the output to the top-down pathway. Owing to the multiple losses in every level, every dilated convolution layer can directly receive the gradient feedback from the loss function. This can ensure that the gradient values at every dilated convolution level in the training process will not be vanished and thus handle the over-fitting problem. This is another important reason that our network can obtain the state-of-art performance of salient object detection, without the need of pretraining on the large dataset (such as ImageNet) like the previous studies.

4. We validated the proposed network architecture on a large range of data sets (six benchmarks). The results demonstrated the effectiveness of this architecture and the superiority to eleven state-of-the-art approaches in the saliency detection task.

2 Methodology

Our architecture can be divided into two parts: bottom-up pathway and top-down pathway. The network architecture is illustrated in Figure 1. The bottom-up pathway aims to extract the low-level and high-level saliency features via the multi-level and pyramid hierarchical structure (see Section 2.1). The top-down pathway intends to recover the high-resolution saliency maps at different semantic degrees by the feature fusion (see Section 2.2). The resulting network thus contains multiple bottom-up and top-down pathway pairs for low-level and high-level feature extraction and fusion. Then we propose a loss function considering three aspects, pixel-level similarity, spatial Euclidean distance and overlapping degree, for comparing the various-level saliency maps with the ground truth. Section 2.3 presents the formulation of the loss function and the network details in the training and implementation.

2.1 Bottom-Up Pathway

We propose a cascade structure with five abstraction levels in the bottom-up pathway. The main component of every level is a block with dilated densely-connected convolutional network (referred as to dilated dense block, DDB), inspired by DenseNet [7]. In contrast to DenseNet, our DDB can enlarge the empirical receptive field, which is the much smaller than the theoretical one in both low-level and high-level layers of DenseNet [32]. It also preserves the resolution of the feature maps, and thus feasible to transfer the feature maps to downstream applications that require spatial detailed information. The input feature maps of all levels except Level 2 and 3 is propagated to the DDB after a convolution layer. In Level 2 and 3, a pooling operator is between the convolution layer and DDB. The channel numbers of output feature maps for all levels are 160, 224, 256, 272, 280, respectively. Besides be the input of the next level, the output feature map of the ith level is convolved by a layer network for producing the highly abstract feature map with 16 channels (denoted by m_i) in the corresponding semantic degree. In the last level, we apply the pyramid pooling module (PPM) [32] to extract the hierarchical global contextual information from m_5, for collecting global context information of the salient object in the largest receptive field of our network. Totally four scales in the pyramid structure are used in the PPM (1×1, 2×2, 3×3 and 6×6). The output of PPM is a feature map with 4 channels (denoted by m_6).

A DDB contains K dilated convolutional layers with dense connectivity. The forward model of the the kth layer in the DDB can be formulated as

$$x_k = H(y_1, ..., y_{k-1}), \quad y_k = D(x_k) \tag{1}$$

where x_k and y_k are the input and ouput of the kth layer, respectively. The function $H(\cdot)$ is the concatenation of the inputs of the preceding $k-1$ layers. The function $D(\cdot)$ is the dilated convolution:

$$y_k(i,j) = \sum_u \sum_v x_k(d \times u, d \times v) \cdot g(i-u, j-v) \tag{2}$$

where $y_k(i,j)$ is the value of the feature map y_k at (i,j), g is the dilated convolutional kernel. u and v are the coordinate offsets in g. For Level 1 to Level 5, the values of the dilated rate d are 1, 1, 2, 4 and 8, respectively.

2.2 Top-Down Pathway

The top-down pathway aims to perceive the saliency maps from the feature maps in different abstraction levels. From the high level to low level, each of the feature maps (from m_6 to m_1) is concatenated with that in the down level successively. The concatenation in the first and second levels additionally requires the upsampling operator. All concatenated feature maps are then separately upsampled after a convolution layer to produce the saliency maps (denoted by $S'_1 \sim S'_5$) with the size 256×256. The value at each pixel within the saliency maps shows the probability that this pixel belongs to the salient object. Then, $S'_1 \sim S'_5$ are concatenated and convolved with a kernel to obtain an extra saliency map integrating the saliency information on all abstraction levels (denoted by S'_0). For $S'_0 \sim S'_5$, we apply the softmax classifier to produce the corresponding salient object images (denoted by $S_0 \sim S_5$) in each level, where every pixel in $S'_0 \sim S'_5$ are determined whether belonging to the salient object. $S_0 \sim S_5$ are compared with the ground truth in training process by the loss function (see Section 2.3). In the testing process,S_0 shows the final result of the salient object detection.

2.3 Training and Implementation Details

As discussed in the introduction part, we utilized multiple loss functions to supervise saliency detector's training procedure. Specifically, we add loss functions at all abstraction levels to independently minimizing difference between S_i and the ground truth, where $i \in \{0,...,5\}$. The supervision loss functions introduce new computed gradients at every level. It is able to propagate the feedback back to every convolution layer in all DDBs for minimizing gradient vanishing problem. At the same time, the introduced multiple predictions compete each other and thus they regularize each other to minimize over-fitting problem. This makes training task becomes easier than without supervisions.

The proposed loss function $L(S)$ for the salient object image S is defined by $L(S) = L_1(S) + L_2(S) + L_3(S)$. L_1 is the pixel-wise weighted cross-entropy [21]:

$$L_1 = -\frac{\sum_{i=1}^{W}\sum_{j=1}^{H}[\lambda w G_{i,j}\log(S_{i,j}) + (1-G_{i,j})\log(1-S_{i,j})]}{W \times H} \tag{3}$$

where G is ground truth. S and G have the same image size $W \times H$, $S_{i,j}$ and $G_{i,j}$ are the pixel value at (i,j) in S and G, respectively. w is a weight to give a balance between saliency region and non-saliency region, defined by $w = (WH - \sum_{i=1}^{W}\sum_{j=1}^{H} S_{i,j})/(\sum_{i=1}^{W}\sum_{j=1}^{H} S_{i,j})$. λ is a parameter to control the influence of w. L_2 is the modified mean absolute errors:

$$L_2 = \sum_{i=1}^{W}\sum_{j=1}^{H} \ln\left(1 + e^{|G_{i,j} - S_{i,j}|}\right) \tag{4}$$

		Amulet	DCL+	DHSNet	DRFI	MSRNet	MST	NLDF	RFCN	SRM	UCF	WSS	Our
DUT-TE	w-F	0.6533	0.6294	0.6991	0.3802	0.6911	0.4596	0.7007	0.5826	0.7146	0.5872	0.5503	0.7349
	max F	0.7783	0.7856	0.8114	0.6497	0.7692	0.5936	0.8123	0.7840	0.8262	0.7710	0.7373	0.8277
	MAE	0.0852	0.0819	0.0655	0.1549	0.0586	0.1630	0.0653	0.0900	0.0588	0.1174	0.1000	0.0610
ECSSD	w-F	0.8413	0.7863	0.8386	0.5191	0.8422	0.6034	0.8393	0.6988	0.8529	0.7885	0.7091	0.8642
	max F	0.9146	0.9003	0.9066	0.7817	0.8889	0.7227	0.9050	0.8904	0.9172	0.9105	0.8556	0.9169
	MAE	0.0592	0.0679	0.0588	0.1704	0.0546	0.1567	0.0626	0.1069	0.0544	0.0778	0.1036	0.0494
HKU-IS	w-F	0.8128	0.7687	0.8140	0.5063	0.8468	0.5865	0.8384	0.6803	0.8353	0.7504	0.7079	0.8496
	max F	0.8954	0.8928	0.8905	0.7771	0.8917	0.7042	0.9020	0.8926	0.9058	0.8858	0.8587	0.9071
	MAE	0.0521	0.0635	0.0524	0.1446	0.0400	0.1389	0.0477	0.0889	0.0459	0.0740	0.0792	0.0420
PASCALS	w-F	0.7547	0.7038	0.7105	0.4560	0.7521	0.5540	0.7268	0.6461	0.7445	0.7125	0.6132	0.7486
	max F	0.8390	0.8166	0.8309	0.6936	0.8404	0.6610	0.8319	0.8350	0.8482	0.8276	0.7807	0.8377
	MAE	0.0993	0.1160	0.0960	0.2112	0.0778	0.1944	0.1007	0.1337	0.0868	0.1274	0.1416	0.0933
SED1	w-F	0.8599	0.7768	0.8696	0.6421	0.8430	0.7359	0.7803	0.7163	0.8139	0.8363	0.7649	0.9027
	max F	0.9219	0.9025	0.9223	0.8699	0.8892	0.8424	0.8885	0.8922	0.9048	0.9216	0.8960	0.9404
	MAE	0.0602	0.0877	0.0528	0.1481	0.0618	0.1238	0.0909	0.1167	0.0753	0.0711	0.1002	0.0404
SED2	w-F	0.8308	0.6462	0.7583	0.6230	0.6549	0.6933	0.6829	0.6417	0.7022	0.7905	0.6830	0.8484
	max F	0.9018	0.8755	0.8800	0.8354	0.7414	0.8001	0.8544	0.8362	0.8616	0.8838	0.8668	0.8985
	MAE	0.0623	0.0925	0.0783	0.1346	0.0972	0.1228	0.1031	0.1132	0.0916	0.0750	0.0982	0.0569
SOD	w-F	0.6767	0.6625	0.6778	0.4294	0.6564	0.5003	0.7040	0.5756	0.6665	0.6382	0.5925	0.7117
	max F	0.7970	0.8255	0.8219	0.6949	0.7793	0.6456	0.8364	0.7899	0.8376	0.7965	0.7740	0.8268
	MAE	0.1414	0.1326	0.1272	0.2231	0.1225	0.2216	0.1244	0.1697	0.1255	0.1640	0.1683	0.1164

Table 1: Comparison between our results and results from 11 state-of-the-art methods on six benchmark data sets in terms of indices: w-F, max F and MAE. The best three scores are shown in red, green, and blue colors, respectively. It is note that "SED1" and "SED2" are two subsets of a dataset "SED". The results show that our approach is at the state of the art.

where the softplus function make this loss function easy to optimize. L_3 is the generalized Dice index [15]:

$$L_3 = 1 - 2\left[\left(w_1 \sum_{i=1}^{W}\sum_{j=1}^{H} S_{i,j}G_{i,j} + w_2 \sum_{i=1}^{W}\sum_{j=1}^{H}(1-S_{i,j})(1-G_{i,j})\right) \Big/ \left(w_1 \sum_{i=1}^{W}\sum_{j=1}^{H}(S_{i,j}+G_{i,j}) + w_2 \sum_{i=1}^{W}\sum_{j=1}^{H}(2-S_{i,j}-G_{i,j})\right)\right] \tag{5}$$

where

$$w_1 = \left(\sum_{i=1}^{W}\sum_{j=1}^{H} G_{i,j}\right)^{-1}, \quad w_2 = \left(\sum_{i=1}^{W}\sum_{j=1}^{H}(1-G_{i,j})\right)^{-1} \tag{6}$$

Because the difference of $S_0 \sim S_5$ with the ground truth are all considered in the network training, the final loss function $\mathcal{L}$ for network training is defined by $\mathcal{L} = \sum_{i=0}^{5} L(S_i)$.

We implement the proposed approach by TensorFlow. In the network training, we use stochastic gradient descent with the momentum 0.9 as the optimization algorithm. The weight decay is 0.0001. The parameter λ is 0.1. The training procedures finished 40 epochs. The learning rate is 0.002 at the initial time, 0.0002 after 25 epochs, and 0.00002 after 30 epochs. All input images are resized to 256×256 for training and testing.

3 Experiments and Results

3.1 Experiment Setup

Datasets. To validate the effectiveness of the proposed network architecture, we carry out comprehensive experiments on six popular benchmark datasets: ECSSD (1000 images) [28],

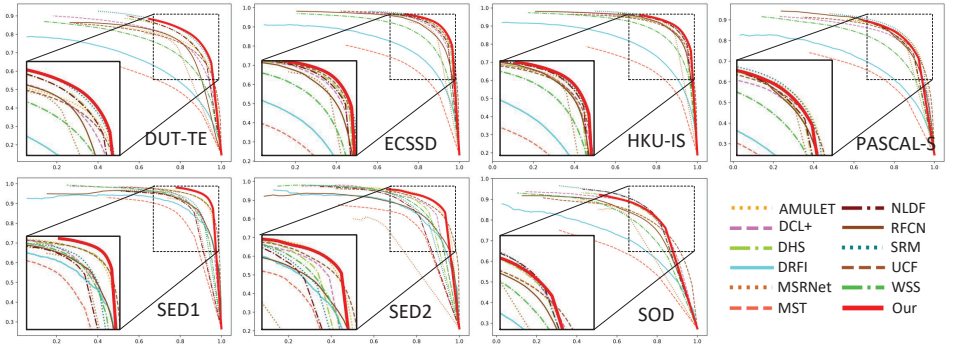

Figure 3: Comparison of precision-recall curves between our approach and the eleven state-of-the-art methods. The x-axis and y-axis correspond to the precision and recall, respectively. The results show that our approach is at the state of the art.

HKU-IS (4447 images) [10], PASCAL-S (850 images) [13], SOD (300 images) [18], DUT-TE (5019 images) [24], SED [1]. SED has two subsets SED1 and SED2. SED1 has 100 images, and each image in SED1 has only one salient object. SED2 has 100 images, each image in SED2 has two salient object. In addition, three public datasets are used for network training, including DUT-TR (10553 images) [24], DUT-OMRON (5168 images) [29], MSRA10K [2] (10000 images).

Evaluation Indices. We evaluated different approaches via comparing ground truth (G) to the results (S). We utilized five evaluation indices: precision, recall, maximum F-measure, weighted F-measure, and mean absolute error in our experiments.

The precision and recall are is defined by

$$Precision = \frac{|S \cap G|}{|S|}, \quad Recall = \frac{|S \cap G|}{|G|} \tag{7}$$

where the operator $|A|$ is to calculate the pixel number within the image region A.

Based on the precision and recall, we formulated the F-measure (F_β) as

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{8}$$

the parameter β gives a balance between the precision and recall, and set $\sqrt{0.3}$ as [26]. The maximum F-measure (max-F) is the maximum value of F-measure, and the weighted F-measure (w-F) is proposed by [17] for handling the existing flaws of F-measure.

The mean absolute error (MAE) is defined by

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |S_{i,j} - G_{i,j}| \tag{9}$$

where W and H are the width and height of the detected salient object image S.

3.2 Comparison with the State-of-the-art Methods

Our approach is compared with eleven state-of-the-art methods, including nine deep learning based methods (Amulet [30], DCL+ [11], DHSNet [14], MSRNet [12], NLDF [16], RFCN

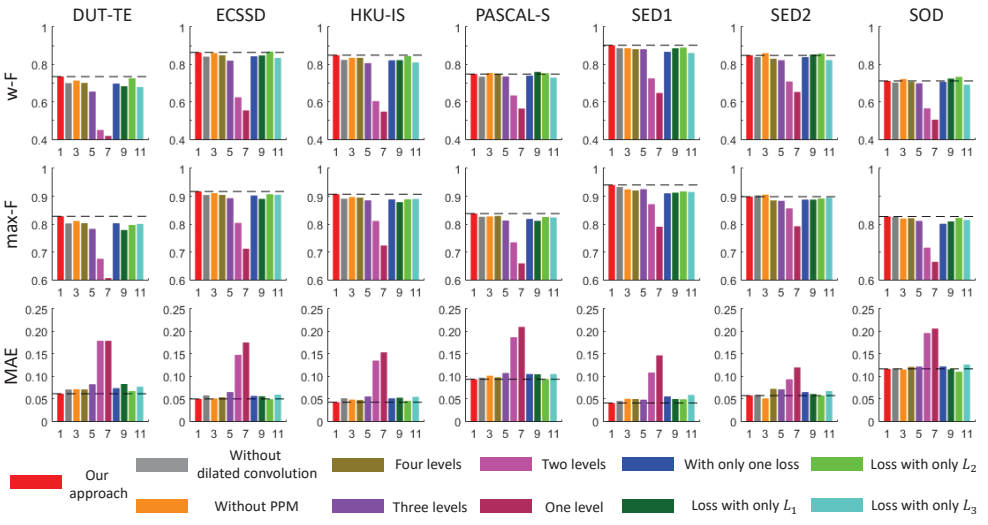

Figure 4: Ablation analysis for comparing the variants of network architecture on the public databases with our approach (red). The black dashed line shows the performance of our approach. "Without dilated convolution" (gray) is to use the convolution kernel with dilated rate of 1 in the DDB, and add two pooling layer after the convolution layer in Level 4 and Level 5, respectively. "Without PPM" (dark orange) is to remove the PPM. "Four levels" to "One level" (gold, purple, magenta and maroon) are to preserve the bottom-up and top-down pathways in different levels (from four levels to one level), and remove the higher-level network structure. "With only one loss" (blue) is remove the loss functions of S_0 to S_4, i.e. only reserve the loss function of S_5. "Loss with only L_1", "Loss with only L_2", "Loss with only L_3" (dark green, green, turquoise) are to reserve the one of L_1, L_2 and L_3 in the loss function $L(S)$, respectively, each of which remove other two components of the loss function. The results show the effectiveness of the network architecture in our approach.

[25], SRM [26], UCF [31], and WSS [24]), and two conventional methods (DRFI [9] and MST [22]) without using deep learning technology. For a fair comparison, we either re-implemented these algorithms with recommended parameter settings or utilized the online source codes provided by the authors.

Table 1 shows the comparison results on the six benchmark data sets. On these data sets, 95.2% indices placed our method on the top three (66.7% first, 9.5% second, 19.0% third). Note that in terms of max-F, our approach ranks at third place except on the PASCAL-S. In terms of MAE, our approach ranks at first place except DUT-TE and PASCAL-S. Overall, our method performs well in these benchmark datasets.

Figure 2 illustrated eight examples of detection results generated from different methods. The results show our approach is able to handle a large range of object's scale changes. Our method is able to draw attention to thin and small regions. This is essential for image understanding and cognitive tasks to perceive small objects. In addition, it is note that our saliency map stick clearly out from background. This may help image understanding tasks make less biased decision from background disturbances.

Figure 3 shows comparison among different methods in terms of precision-recall curves. It indicates that our approach is among the top contenders.

3.3 Ablation Analysis

The ablation analysis aims to investigate the effectiveness of our network architecture. Figure 4 illustrate the comparison of different variants with our approach (red bars). The comparison with the gray bars shows the effectiveness of the dilated convolution by comparing the original DenseNet structure, i.e. set the the dilated rate as 1 and add two pooling layers after the convolution layers in Level 4 and Level 5, respectively. To analyze the relative contributions of different levels and the PPM of our approach, a comprehensive comparison of their performance are evaluated on the state-of-the-art methods (see the dark orange, gold, purple, magenta and maroon bars). Then we apply only one loss function to evaluate the difference between S_5 and the ground truth in the training process, i.e. to neglect the all lower-level pathways and only keep Level 5. The results show that plugging in multiple loss functions at the different levels can facilitate the performance improvement (see the blue bars). Finally, we show that the combination of the three components in our loss function is effective because the performance will decrease when using one of the three components alone (see the dark green, green and turquoise bars).

4 Discussion

The main superiority of our network is to design a backbone network for SOD task. The backbone network has a special architecture to preserve the spatial detail information and capture sufficient global semantic information. Most previous SOD methods are based on the existing classification networks like VGG and ResNet. Their architectures suit for the classification task but not suit for the SOD task, because the classification task only needs that the network has enough effective receptive field to extract sufficient global semantic information. In constrast, the SOD task requires that the network can preserve the spatial detail information besides the effective receptive field to predict accurate the salient object boundaries. In order to preserve the spatial detail information, our network includes two schemes. First, we just use two pooling layers to decrease the resolution of feature maps for storing more channels of feature maps in the limited GPU memory. In the other scheme, we use the holistic deep feature maps from low to high level to infer the final saliency map. For acquiring enough effective receptive field, we apply the dilated convolution with varied dilated rate to obtain bigger receive field than traditional convolution, and employ PPM to efficient capture the global semantic information. In addition, multiple arbitrary losses are introduced to optimize our network in order to train from scratch.

5 Conclusion

In this paper, we develop a novel end-to-end holistic and deep pyramid neural network approach for saliency detection. The proposed multi-level pyramidal hierarchical architecture facilitates to effectively extracting and fusing the high-level semantic information and low-level fine-grain information to produce high-resolution saliency maps. Deep supervision introduced extra gradients to void gradient vanishing problem. Multiple prediction tasks regularize each other to minimize over-fitting problem. This architecture allows us to train the network from scratch. Extensive experimental results demonstrate the performance improvement of our approach in the saliency detection task on six benchmark datasets comparing to eleven state-of-the-art methods.

References

[1] Ali Borji. What is a salient object? a dataset and a baseline model for salient object detection. *IEEE Transactions on Image Processing*, 24(2):742–756, 2015.

[2] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[3] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163(10):90–100, 2017.

[4] Keren Fu, Chen Gong, Yixiao Yun, Yijun Li, Irene Yu-Hua Gu, Jie Yang, and Jingyi Yu. Adaptive multi-level region merging for salient object detection. In *The British Machine Vision Conference (BMVC)*, 2014.

[5] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H. S. Torr. Deeply supervised salient object detection with short connections. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3203–3212, 2017.

[6] Ping Hu, Bing Shuai, Jun Liu, and Gang Wang. Deep level sets for salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2300–2309, 2017.

[7] Gao Huang, Zhuang Liu, and Laurens van der Maaten. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[8] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Tie Liu, and Nanning Zheng. Automatic salient object segmentation based on context and shape prior. In *The British Machine Vision Conference (BMVC)*, 2011.

[9] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: a discriminative regional feature integration approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2083–2090, 2013.

[10] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5455–5463, 2015.

[11] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 478–487, 2016.

[12] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 247–256, 2017.

[13] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 280–287, 2014.

[14] Nian Liu and Junwei Han. DHSNet: deep hierarchical saliency network for salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 678–686, 2016.

[15] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):353–367, 2011.

[16] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6593–6601, 2017.

[17] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2014.

[18] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE Conference on Computer Vision (ICCV)*, pages 416–423, 2001.

[19] Tam V. Nguyen and Luoqi Liu. Salient object detection with semantic priors. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4499–4505, 2017.

[20] Qinmu Peng, Yiu ming Cheung, Xinge You, and Yuan Yan Tang. A hybrid of local and global saliencies for detecting image salient region and appearance. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(1):86–97, 2017.

[21] 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 (MICCAI)*, pages 234–241, 2015.

[22] Wei Chih Tu, Shengfeng He, Qingxiong Yang, and Shao Yi Chien. Real-time salient object detection with a minimum spanning tree. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2334–2342, 2016.

[23] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3183–3192, 2015.

[24] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3796–3805, 2017.

[25] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 825–841, 2016.

[26] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stage-wise refinement model for detecting salient objects in images. In *IEEE Conference on Computer Vision (ICCV)*, pages 4039–4048, 2017.

[27] Jeremy M. Wolfe and Todd S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5:1–7, 2004.

[28] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1162, 2013.

[29] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3166–3173, 2013.

[30] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: aggregating multi-level convolutional features for salient object detection. In *IEEE Conference on Computer Vision (ICCV)*, pages 202–211, 2017.

[31] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *IEEE Conference on Computer Vision (ICCV)*, pages 212–221, 2017.

[32] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2017.

[33] Wenbin Zou, Kidiyo Kpalma, Zhi Liu, and joseph Ronsin. Segmentation driven low-rank matrix recovery for saliency detection. In *The British Machine Vision Conference (BMVC)*, 2013.