Regularized Densely-connected Pyramid Network for Salient Instance Segmentation

Yu-Huan Wu, Yun Liu, Le Zhang, Wang Gao, and Ming-Ming Cheng, *Senior Member, IEEE*

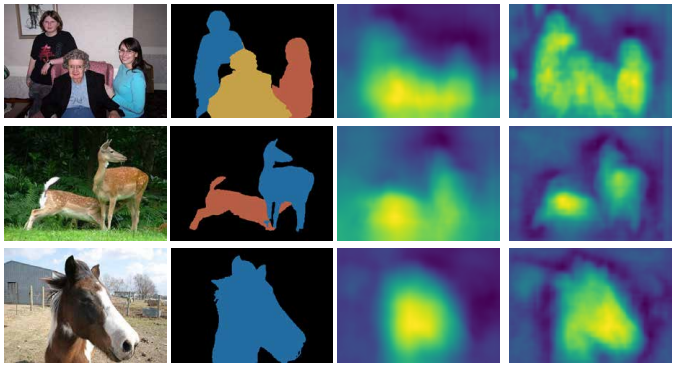
***Abstract*—Much of the recent efforts on salient object detection (SOD) have been devoted to producing accurate saliency maps without being aware of their instance labels. To this end, we propose a new pipeline for end-to-end salient instance segmentation (SIS) that predicts a class-agnostic mask for each detected salient instance. To better use the rich feature hierarchies in deep networks and enhance the side predictions, we propose the regularized dense connections, which attentively promote informative features and suppress non-informative ones from all feature pyramids. A novel multi-level RoIAlign based decoder is introduced to adaptively aggregate multi-level features for better mask predictions. Such strategies can be well-encapsulated into the Mask R-CNN pipeline. Extensive experiments on popular benchmarks demonstrate that our design significantly outperforms existing state-of-the-art competitors by 6.3% (58.6% vs. 52.3%) in terms of the AP metric. The code is available at https://github.com/yuhuan-wu/RDPNet.**

***Index Terms*—salient instance segmentation, feature pyramid, RoIAlign**

I. Introduction

As a fundamental image understanding technique, salient object detection (SOD) aims at segmenting the most eye-attracting objects in a natural image. Although recent SOD approaches [1]–[5] have achieved much success, their generated saliency maps cannot discriminate different salient instances, which has prevented many applications from applying SOD for instance-level image understanding [6]. Motivated by [7], in this paper, we tackle the more challenging case of SOD, called salient instance segmentation (SIS). SIS segments salient objects from an image and discriminates salient instances by associating each instance with a different label. SIS can facilitate more advanced tasks than SOD, such as image captioning [8], weakly-supervised semantic/instance segmentation [9], [10], and visual tracking [11].

The MSRNet [7] made the first attempt to detect salient instances by adopting several isolated processing steps. However, its performance was usually limited in challenging scenarios because it was not end-to-end trainable. The S4Net [12] replaced *RoIAlign* in Mask R-CNN [13] with the *RoIMasking*



(a) (b) (c) (d)

Figure 1. Visualizations for the feature maps after passing FPN and our proposed regularized densely-connected pyramid (RDP). (a) Source images; (b) Corresponding ground truth; (c) Visualized maps for the feature maps after FPN; (d) Visualized maps for the feature maps after the proposed RDP. The visualized feature maps directly obtained by the FPN look coarser and hard to recognize objects. With our proposed RDP, it is much easier to recognize each salient instance's locations and shapes.

to keep the scale of the feature maps and leverage the nearby background of objects. Although much better performances were reported, it was far from satisfactory because only a limited feature level was utilized to decode salient instances. One may argue that a natural solution is to employ the Feature Pyramid Network (FPN) [14] and solve this task using the feature pyramid. FPN builds the feature pyramid via the top-down pathway and lateral connections from the backbone. With this network, small and large objects are more likely to be detected in the pyramid's low and high levels, respectively. Therefore, apart from detecting the salient objects, much of the information flow was devoted to detecting the small and unnoticeable objects with the top-down pathway. Naively applying the FPN architecture for SIS is suboptimal because salient objects are often much larger and distinctive than noisy background and uninteresting objects.

Motivated by this, we focus on enhancing the side predictions by providing each side branch with richer feature hierarchies from deep networks to locate the object and recover its details. We achieve this by proposing the regularized densely-connected pyramid (RDP) network, which provides richer feature hierarchies for each branch with dense connections. In this way, each level can leverage both high-level semantic and low-level fine-grained features. However, directly leveraging dense connections may yield noisy predictions due to different receptive fields of features with different feature levels. To this end, we propose to regularize such dense connections using the attention mechanism to promote informative features and

Y.-H. Wu and M.M. Cheng are with the TKLNDST, College of Computer Science, Nankai University, Tianjin 300350, China. (E-mail: wuyuhuan@mail.nankai.edu.cn, cmm@nankai.edu.cn)

Y. Liu is with ETH Zurich. (E-mail: yun.liu@vision.ee.ethz.ch)

L. Zhang is with the School of Information and Communication Engineering, University of Electronic Science and Technology of China. (E-mail: zhangleuestc@gmail.com)

W. Gao is with the Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, Beijing, China. (E-mail: gaowang_fly@163.com)

M.-M. Cheng (cmm@nankai.edu.cn) and W. Gao (gaowang_fly@163.com) are corresponding authors.

suppress non-informative ones from all feature levels of the feature pyramid.

Our effort starts with Mask R-CNN [13] that first detects bounding boxes and then adopts *RoIAlign* to predict the binary mask for each region of interest (RoI). Specifically, we propose the regularized densely-connected pyramid (RDP) network mentioned above to better enhance the feature pyramid with different scales while keeping semantic features for detecting salient instances. More specifically, each level of features will be fused with not only its successive bottom features, as done in other works [12], [14]–[17], but also features from all the lower levels. The RDP network only costs 0.7ms, which can be ignored in affecting the speed of the whole network. Fig. 1 shows the superiority of RDP in feature learning compared with the FPN. Besides, for mask prediction, traditional strategies like Mask R-CNN only use a specific feature level. Which level is used is determined by the size of objects. This design is suboptimal for SIS, and leveraging all feature levels is a better strategy. Motivated by this, we propose to leverage the feature maps from all feature levels with a novel multi-level *RoIAlign* operation for extracting hierarchical RoIs for better mask prediction. Extensive experiments demonstrate that the proposed method achieves state-of-the-art performance and far surpasses previous competitors in terms of all metrics. With an NVIDIA TITIAN Xp GPU, the proposed method runs at 45.0fps for images of the $\sim 320 \times 480$ size and is thus suitable for real-time applications.

Overall, our main contributions are summarized as below:

- We propose regularized dense connections to attentively promote informative features and suppress non-informative ones at each stage of the feature pyramid, providing richer bottom-up information flows.
- We further propose a novel multi-level RoIAlign based decoder to pool multi-level features for better mask predictions adaptively.
- We empirically evaluate the proposed method on two popular SIS datasets and demonstrate its superior accuracy and better efficiency.

II. Related Work

A. Salient Object Detection

SOD aims to detect salient objects or regions in natural images. Conventional methods [2], [18]–[21] mainly focus on designing hand-crafted features and better prior strategies for SOD. Later, some learning-based features [2] were studied as well. Due to their limited representational ability, these methods have been suppressed by the deep learning-based methods. Motivated by the success of convolutional neural networks (CNNs) and fully convolutional networks (FCNs) [22], many FCN-based SOD networks were proposed [1], [3]–[5], [23]–[31]. For example, Wang *et al.* [3] developed a recurrent FCN architecture for saliency prediction. Liu *et al.* [23] presented a deep hierarchical saliency network to learn a coarse global prediction and refine it hierarchically and progressively by integrating local information. Inspired by [32], [33], Hou *et al.* [24] introduced short connections for side-outputs to enrich multi-scale features. Zhang *et al.* [1] introduced a bi-directional structure to adaptively aggregate multi-level features. Wang *et al.* [34] proposed to detect salient objects globally and recurrently refine the saliency maps. Liu *et al.* [35] proposed a pixel-wise contextual attention network to selectively attend to each pixel's informative context locations. Liu *et al.* [4] proposed various pooling-based modules to strengthen the feature representations with real-time speed. More details of the development in SOD can refer to [36]–[39]. Although these methods can detect saliency maps accurately, they cannot discriminate different salient object instances.

B. Instance Segmentation

Similar to object detection, early instance segmentation works [40]–[42] focus on classifying segmented proposals generated by object proposal methods [43]–[46]. Li *et al.* [47] first proposed an end-to-end fully convolutional instance segmentation (FCIS) framework. He *et al.* [13] extended Faster R-CNN [48] to Mask R-CNN by replacing *RoIPool* with *RoIAlign* for more accurate RoI generation. They added a parallel mask head with the box head in Faster R-CNN for mask prediction using the feature pyramid's RoI features. Mask scoring (MS) R-CNN [49] combines the mask confidence score and the localization score and is thus more precise for scoring the detected instances. HTC [50] proposes a hybrid multi-stage cascade for both box and mask detection. Based on FCOS [51], CenterMask [52] designs spatial-attention-guided mask prediction for anchor-free instance segmentation. BlendMask [53] achieves instance segmentation via a blender with the learned bases and instance attentions. SOLO [54] segments object for each location. DetectoRS [55] proposes the recursive feature pyramid and switchable atrous convolution for better performance.

C. Feature Pyramid Enhancement

The feature pyramid is known as a powerful tool for strengthening multi-scale feature representations [56]. The necessity of feature pyramid enhancement has also been demonstrated in detecting locations [14] or segmenting objects [57]. The early successor FPN [14] builds the feature pyramid via the top-down pathway and lateral connections from the backbone feature pyramid. PANet [58] builds upon FPN and adds an extra bottom-up path augmentation. NAS-FPN [59] extends the idea of FPN by learning the scalable feature pyramid architecture using neural architecture search (NAS). EfficientDet [60] proposes BiFPN, which optimizes multi-scale feature fusion in a more efficient bidirectional manner.

D. Salient Instance Segmentation

SIS is a relatively new problem that shares similar spirits with both SOD and instance segmentation. It is more challenging than SOD because it segments salient objects and meanwhile differentiates different salient instances. One possible solution is to derive the salient instances directly from the saliency map using some post-processing techniques. For example, Li *et al.* [7] proposed a two-stage solution, called MSRNet, which first produces saliency maps and salient object contours

that are then integrated with MCG [44] for SIS. Although MSRNet can learn from the saliency maps, as the two stages are optimized isolatedly, the results are not satisfactory. To overcome the isolated optimization difficulties, recently, Fan *et al.* [12] introduced an end-to-end single-stage framework based on the Mask R-CNN [13]. They learned to mimic the strategy of GrabCut [61] and used the so-called *RoIMasking* to incorporate foreground/background separation explicitly. They also designed a customized segmentation head with dilated convolutions to retrieve instance masks from the coarsest feature level. Instead of using a single specific feature level with limited semantic features as done in existing methods, we propose to use the regularized densely-connected pyramid (RDP) networks to extract richer feature hierarchies with higher contrasts (as in Fig. 1) from all feature levels. Our design significantly releases the burden of accurately detecting salient instances and retrieving binary masks for each salient instance.

III. Our Approach

A. Feature Pyramid Enhancement

The feature pyramid, which is usually understood as a group of feature maps with different resolutions, has demonstrated its superiority in various computer vision tasks. One notable application is object detection, which aims to detect semantic objects' locations accurately. As there are large-scale variations for natural objects, directly detecting targets' accurate locations by simply using features from one scale is extremely challenging. Therefore, many researchers attempt to detect semantic objects with the feature pyramid. Our method naturally belongs to this family. We propose a densely-connected pyramid (DP) network and the advanced regularized densely-connected pyramid (RDP) network for the feature pyramid enhancement. We elaborate on the main idea below.

1) Problem Formulation: Given an image as the input and a base network (*e.g.*, ResNet [62]) for feature extraction, we can first derive a set of side-outputs from multiple stages in this network. Assume that we have access to multiple scales of features $\{C_m, C_{m+1}, \cdots, C_k\}$ from the m-th to k-th stage, corresponding to the finest and coarsest feature maps. Typically, m will be 2 as defined in two-stage detectors like Faster R-CNN [14], [48] or 3 as defined in one-stage detectors like RetinaNet [15]. k is typically 5 as defined in both kinds of detectors [14], [15], [48].

2) The Top-down Style: In order to leverage both high-level semantics and low-level fine details as mentioned above, the well-known FPN [14] proposes a top-down architecture with lateral connections to strengthen the capacity and representability of each side-output. Such a strategy has been demonstrated very powerful especially for detecting small and tiny objects and has been extensively used in many other approaches. Suppose that the feature pyramid enhanced by FPN is called $P = \{P_m, P_{m+1}, ..., P_k\}$. This enhancement operation can be formulated as:

$$P_k = \mathcal{F}[\phi(C_k)], \tag{1}$$

$$P_i = \mathcal{F}[\phi(C_i) + Upsample(P_{i+1})], m \leq i < k, \tag{2}$$

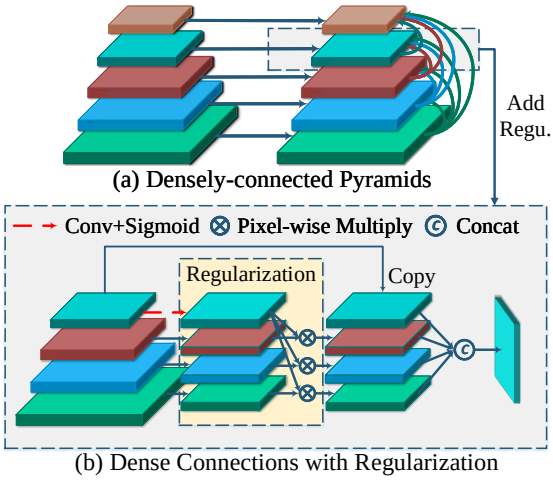


Figure 2. Illustration of the proposed regularized densely-connected pyramid (RDP) network for feature pyramid enhancement. (a) The densely-connected pyramid (DP) network; (b) Dense connections with regularization. For simplicity, we illustrate the regularization with only the 4^{th} feature level. RDP is DP with the regularization at each feature level.

where ϕ represents a 1×1 convolution layer to reduce the channels of C_i. $\mathcal{F}$ represents the feature fusion module which consists of a single 3×3 convolution layer. The upsampling factor for P_{i+1} is 2 and we use the bilinear interpolation for upsampling. For the coarsest feature map C_k, this enhancement operation is simplified done by passing a single 3×3 convolution.

Such a strategy, however, is suboptimal for SIS. Recall that the objective of this task is to detect salient instances and ignore other non-salient ones that usually have a relatively smaller size. In Equ. (2), each side branch only has limited bottom-up information because it only leverages the features of two successive layers. In this way, higher levels in the pyramid have limited access to the low-level fine-grained details and thus may fail to recover the instance boundaries. Similarly, the lower levels in the pyramid lack the high-level semantic information and thus may not be good at accurately locating the salient objects and identifying their instance labels. To address this problem, we provide our solution below.

3) The Bottom-up Densely-connected Pyramid Network: A straightforward solution to overcome the above-mentioned disadvantages of FPN, as proposed in [58], is to build a progressive bottom-up lateral connection and recreate a new feature pyramid:

$$P'_m = \mathcal{F}(P_m), \tag{3}$$

$$P'_{j+1} = \mathcal{F}[P_{j+1} + Downsample(P'_j)], m < j \leq k-1, \tag{4}$$

where P'_k is the re-generated feature map of the new feature pyramid. This solution naturally follows a progressive manner of the FPN and is applied in instance segmentation [58]. We take inspiration from this architecture and make necessary amendments. For each feature level in the network, instead of only merging two successive levels, we merge features from many other levels as well. This design is advantageous because each stage is given a much richer information flow from all

its bottom layers. More specifically, we achieve this by adding dense connections, which can be formulated as

$$P'_j = \mathcal{F}\{\phi[Concat(P'_m, P'_{m+1}, ..., P'_{j-1}, P_j)]\}, \quad (5)$$

where we have $m < j \leq k$ and m represents the index of the first stage of the feature pyramid. In the concatenation operation, feature maps $P'_m, P'_{m+1}, ..., P'_{j-1}$ are all downsampled to the size of P_j. We use the 1×1 convolution operation ϕ to reduce the channels to that of P_j.

4) Regularized Densely-Connected Pyramid Network: The bottom-up dense connections essentially expand the input space for each side branch. However, as features from different layers usually have different receptive fields, they are usually not very compatible in discovering the fine details of the object due to the scale conflict. To this end, we further regularize the dense connections with the well-established self-attention mechanism. To compute the new feature maps P'_j, we first create spatial regularization based on the feature map P_j of the current scale:

$$R_j = \sigma\mathcal{F}(P_j), \quad (6)$$

where R_j is the attention map for the regularization and σ denotes the sigmoid function for each pixel. By reducing the effect of the scale conflict during the feature concatenation, we apply this regularization to the feature maps with identical attention maps R_j in the feature fusion except for P_j:

$$P^r_t = R_j \otimes Downsample(P'_t), m \leq t < j, \quad (7)$$

where P^r_t is the regularized feature map from other scales. We perform the downsampling operation to features maps from other scales of the same size as P_j. The symbol $\otimes$ denotes the element-wise multiplication. Overall, the regularized dense connections for enhancing the feature pyramid can be formulated as

$$P'_j = F[Concat(P^r_m, P^r_{m+1}, ..., P^r_{j-1}, P_j)], m < j \leq k. \quad (8)$$

We provide an illustration of the proposed RDP in Fig. 2 for better understanding.

B. Multi-level RoIAlign for Mask Prediction

Mask prediction is essential for SIS as it directly determines the accuracy of the mask for each salient instance. As shown in Fig. 3 (c), Mask R-CNN [13] uses a specific feature level, which depends on the size of the object of interest, for mask prediction using *RoIAlign*. Although this option of determining which feature level is used in *RoIAlign* can adaptively extract masks for objects of different sizes, this is suboptimal for SIS and a better strategy is to leverage all the feature levels. More specifically, we propose an efficient yet well-performing multi-level *RoIAlign* with a decoder to leverage all feature levels. Fig. 3 (d) illustrates our idea. After the multi-level *RoIAlign* layer, we derive a tiny feature pyramid specifically for the mask prediction. The next decoder is to progressively decode the binary masks from the tiny feature pyramid. The decoder consists of the lateral connections and some feature fusion operations. Since the strides of the top two feature maps are very large, they are *RoIAligned* to the same size of RoIs, and we perform element-wise sum for these two RoIs. Other feature maps are *RoIAligned* to different sizes of RoIs.

With this decoder, we first use *RoIAlign* to adaptively align features from all levels and then retrieve binary masks based on the aligned features. For the feature fusion between two adjacent feature maps of different sizes, we first perform bilinear interpolation to upsample them to the size of the finer feature map by a factor of 2. Then, we use the element-wise sum to fuse these two feature maps and add a 3×3 convolution layer to generate the new feature maps for the next feature fusion. Finally, we get the finest feature maps, on which we perform a 1×1 convolution to predict the binary masks.

C. Overall Pipeline

The regularized densely-connected pyramid and the multi-level *RoIAlign* layer are encapsulated into a Mask R-CNN based pipeline, as displayed in Fig. 3. The functionality of each component is presented in the following.

1) Feature Extraction: We adopt the widely used ResNet [62] as our backbone network, which has been pretrained on the ImageNet dataset [63]. The base feature pyramid follows the architecture of FPN [14]. Since we use the one-stage detector [51] for box regression, we follow [51] to generate two extra feature maps, P_6 and P_7, by connecting two 3×3 convolutions with a stride of 2 after P_5. P_6 and P_7 are added to the feature pyramid, so the feature pyramid after passing FPN is $\{P_3, P_4, P_5, P_6, P_7\}$. All feature maps in this feature pyramid are with 256 channels. Then, we build the regularized densely-connected pyramid (RDP) from P'_3 to P'_7, as introduced in Section III-A4 and Fig. 2. The number of output channels is still 256 for all feature maps in the reconstructed feature pyramid. Fig. 1 displays the visualization of feature maps after passing FPN and our proposed RDP. We find that although feature maps derived by FPN have captured the locations of salient instances, the activation or high responses are very coarse or cannot even recognize the number of salient instances in each image. In contrast, the feature maps from our proposed RDP network have more precise activation and can help the base detector better to detect the bounding box of each salient instance. This design further enhances the mask head towards obtaining better masks for the detected salient instances.

2) Box Regression: To quickly detect the salient instances, we do not apply a heavy two-stage detector that contains an RPN [48] head to generate object proposals and classifies these object proposals with the box head because it is too slow for SIS. Instead, we use the one-stage detector FCOS [51] as our base detector. This detector consists of four convolution-ReLU layers with 256 channels, and the box regression is performed at each feature level with this shared-parameters head. The details for calculating the box proposals from the final feature map can refer to [51]. In this part, we will derive many box proposals with their confidence scores in each feature level. We concatenate them and leave the top 1000 boxes with a confidence score larger than 0.05. After that, a non-maximum suppression (NMS) operation is conducted on the boxes and then keep at most top 100 boxes for predicting their corresponding binary masks.

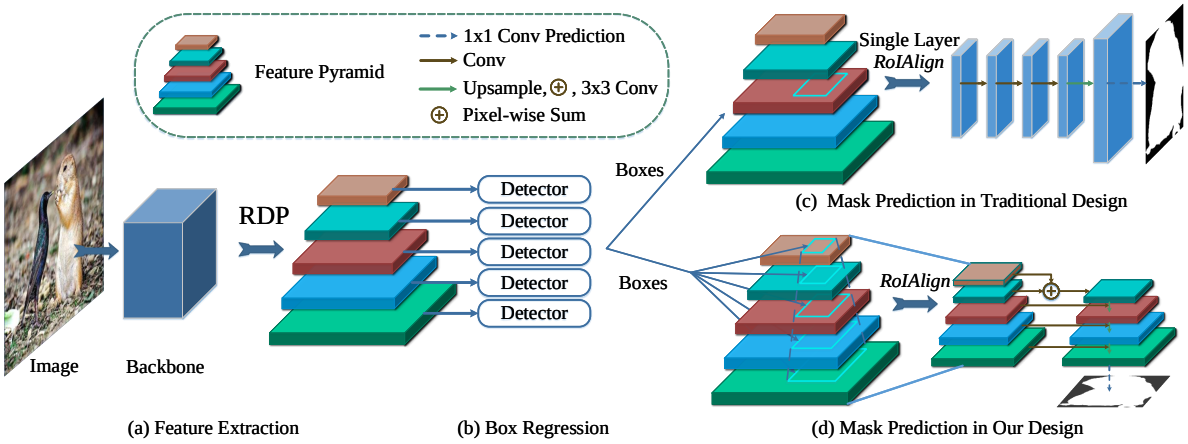


Figure 3. The overall pipeline of the proposed method. (a) In the feature extraction part, RDP is the regularized densely-connected pyramid network, as illustrated in Fig. 2. (b) We use the base detector [51] for box regression at each feature level. (c) The traditional design for mask prediction only uses a single layer to decode the binary masks. (d) Our design for mask prediction uses all feature levels to decode binary masks by a simple decoder.

3) Mask Prediction: In the box regression, we detect the salient instances in the box level. Since our final goal is to predict the instance-level segmentation, mask prediction is necessary to retrieve the corresponding binary mask for each salient instance. We make a further improvement to Mask R-CNN by leveraging the feature maps of all feature levels ($\{P'_3, P'_4, P'_5, P'_6, P'_7\}$) for retrieving binary masks for salient instances. After the multi-level *RoIAlign* layer, the sizes of the feature maps $\{D_3, D_4, D_5, D_6, D_7\}$ are displayed in Table I. Please refer to Section III-B for the implementation of the decoder. After passing this decoder, we use a simple 1×1 convolution layer to predict the final masks for the detected salient instances.

Table I
FEATURE MAP SIZE FOR EACH CHANNEL AFTER THE MULTI-LEVEL *RoIAlign* LAYERS. SINCE P'_7 AND P'_6 ARE VERY SMALL, D_7 AND D_6 ARE SAMPLED WITH THE SAME SIZE. THE SIZE OF THE FINAL MASK FOR EACH SALIENT INSTANCE IS 32×32.

Name	D_7	D_6	D_5	D_4	D_3
Size	4×4	4×4	8×8	16×16	32×32

4) The Loss Function: Our pipeline has two key parts that need supervisions: box regression and mask prediction. A foreground box classification loss L_{cls} and a coordinate regression loss L_{reg} are applied in the box regression branch. Note that L_{cls} is the focal loss [15] and L_{reg} is the IoU loss proposed in [64]. To further get rid of the bad effect of too many low-quality boxes, we apply the centerness loss L_{center} proposed in [51] to ignore the boxes whose centers are far away from the centers of salient instances. For mask prediction, we use the standard cross-entropy loss as the mask loss L_{mask}. Hence we obtain the final loss $L = L_{cls} + L_{reg} + L_{center} + L_{mask}$ to supervise the whole network.

IV. EXPERIMENTS

In this section, we will first introduce the datasets and evaluation metrics used in our experiments, as in Section IV-A. Implementation details will be described in Section IV-B. We will carefully examine our proposed designs and demonstrate their effectiveness in Section IV-C. The results of our method and the comparison with previous state-of-the-art methods will be provided in Section IV-D.

A. Dataset and Evaluation Metric

1) Datasets: We adopt two popular datasets in our experiments, *i.e.*, ISOD and SOC datasets. The ISOD dataset is proposed by Li *et al.* [7]. It contains 1000 images with salient instance annotations. We follow the previous work [12] to use 500 images for training, 200 images for validation, and another 300 images for testing. The SOC dataset is proposed by Fan *et al.* [65]. This dataset consists of 3000 images in cluttered scenes with salient instance annotations. Among them, 2400 images are used for training and the other 600 images are used for testing.

2) Evaluation Metrics: Previous works use the mAP metric with a specific threshold such as 0.5 (standard) or 0.7 (strict) to determine whether a detected instance is a true positive (TP), similar to the evaluation in the PASCAL VOC challenge [66]. However, as this metric is not enough to fully reflect the quality of detectors, the MS-COCO evaluation metric [67] has been widely used in mainstream object detection and instance segmentation. We follow the MS-COCO evaluation metric [67] to use mAP@$\{0.5{:}0.05{:}0.95\}$ as the primary metric, since it can better reflect the detection quality. We also report mAP@0.5 and mAP@0.7 for reference, as done in related works [13]–[15], [49], [68]. For simplicity, we use "AP", "AP_{50}", and "AP_{70}" to stand for mAP@$\{0.5{:}0.05{:}0.95\}$, mAP@0.5, and mAP@0.7, respectively.

Table II
EVALUATION ON THE ISOD VALIDATION SET FOR VARIOUS DESIGN CHOICES. THE FIRST LINE REFERS TO THE BASELINE OF FPN. NP IS THE NATURAL PROGRESSIVE BOTTOM-UP STYLE FOR BUILDING THE NEW FEATURE PYRAMID. DP DENOTES THE PROPOSED METHOD THAT REBUILDS THE FEATURE PYRAMID WITH DENSE CONNECTIONS. RDP MEANS TO ADD THE PROPOSED REGULARIZATION TO DP. MRA REPRESENTS THE PROPOSED MULTI-LEVEL *RoIAlign*.

DP	RDP	MRA	AP	AP_{50}	AP_{70}
-	-	-	54.2%	83.3%	69.7%
✔			55.1%	84.4%	71.0%
		✔	56.3%	85.5%	71.2%
	✔		56.4%	85.4%	72.0%
	✔	✔	**57.4%**	**86.1%**	**73.8%**

B. Implementation Details

In this paper, we use the popular PyTorch [69] and Jittor [70] framework to implement our method. If not specially mentioned, we apply the widely used ResNet-50 [62] as the backbone network. In the network training, maybe there is no box satisfying the threshold of the confidence score for NMS, especially in the early training stage, so we add the ground-truth boxes to the results of detected salient instances in the training to prevent such a situation to take place. We only use horizontal flipping as the data augmentation, and each input image is resized as the shorter side is 320 pixels and the longer side follows the initial image aspect ratio but is limited to a maximum value of 480 pixels. We use a single NVIDIA TITAN Xp GPU for all experiments. We use the SGD optimizer with the weight decay of 10^{-4} and the momentum of 0.9. Each mini-batch contains four images. The initial learning rate is 0.0025. For the ISOD dataset [7], the learning rate is divided by 10 after 6K iterations, and we train our network for 9K iterations in total. For the SOC dataset [65], the learning rate is divided by 10 after 24K iterations, and we train our network for 36K iterations in total. Due to the small batch size, all the BatchNorm layers of the backbone network are frozen during training. The 3×3 convolution layers of the box regression head and mask prediction head are with the group normalization [71]. The number of output channels of each 3×3 convolution layer is 128 in the mask prediction head.

C. Ablation Study

In this part, we evaluate the effect of various designs on the ISOD dataset. We use its training set for training and report results on its validation set. If not mentioned, we use the ResNet-50 as the backbone for our network.

1) Effect of DP and RDP: As mentioned in Section III-A4, we propose to create the RDP to fill the vacancy of the FPN. Here, we view FPN as our baseline and evaluate four design choices: i) NP, *i.e.*, the naive progressive bottom-up style for building the new feature pyramid; ii) DP, *i.e.*, the proposed method that rebuilds the feature pyramid with dense connections; iii) RDP, *i.e.*, adding the proposed regularization to the dense connections in DP; iv) MRA, *i.e.*, the proposed multi-level *RoIAlign*. Table II shows the evaluation results on the ISOD validation set. If we add DP without regularization, the metric of AP will be improved by 0.9% compared with FPN. When we add the regularization to DP, a relative 1.3% improvement over DP is observed, indicating that regularization is vital for the proposed densely-connected pyramid. Note that the proposed RDP is very efficient and only costs 0.7ms for a 320×480 input image, making it have little effect on the speed of the whole network.

Table III
EVALUATION ON THE ISOD VALIDATION SET FOR PARTIALLY APPLYING DP/RDP TO A PART OF SIDE-OUTPUTS. $P_3 \sim P_5$ MEANS FROM P_3 TO P_5. $P_3 \sim P_7$ MEANS ALL SIDE-OUTPUTS IN THE FEATURE PYRAMID.

Side-outputs	DP	RDP	AP	AP_{50}	AP_{70}
-	-	-	54.2%	83.3%	69.7%
$P_3 \sim P_5$	✔		54.4%	83.7%	70.3%
$P_3 \sim P_7$	✔		55.1%	84.4%	71.0%
$P_3 \sim P_5$		✔	55.8%	84.9%	71.3%
$P_3 \sim P_7$		✔	56.4%	85.4%	72.0%

2) Effect of Multi-level RoIAlign: The existing research usually predicts object masks using the mask head proposed by Mask R-CNN [13], which predicts masks from a specific feature level. Instead, we propose a top-down progressive mask decoder to utilize all feature levels for object mask prediction, namely multi-level *RoIAlign* (MRA). The comparison between MRA and the traditional *RoIAlign* can be found in Table II. One can see that applying MRA on the baseline brings 2.1%, 2.2%, and 1.5% improvement in terms of AP, AP_{50}, and AP_{70}, respectively. We can also observe that the introduction of MRA based on the baseline with the RDP further leads to an improvement of 1.0%, 0.7%, and 1.8% in terms of AP, AP_{50}, and AP_{70}, respectively. This demonstrates the significance of the proposed MRA in accurate mask prediction by leveraging all feature levels. Overall, the proposed method achieves 3.2% higher AP, 2.8% higher AP_{50}, and 4.1% higher AP_{75} than the baseline of FPN.

3) Partially Applying DP and RDP: Our initial design considers all feature levels (P_3 - P_7) to reconstruct the feature pyramid. Among them, the top 2 feature levels (P_6 and P_7) are generated from P_5 using only two 3×3 convolutions. In this section, we further evaluate the effectiveness of DP and RDP by applying them to a part of side-outputs. Specifically, we only apply DP/RDP to three side-outputs, *i.e.*, P_3, P_4, and P_5, excluding P_6 and P_7. The experimental results are shown in Table III. We could see that applying DP/RDP to only three side-outputs performs better than the baseline, but performs worse than applying DP/RDP to five side-outputs, indicating that DP/RDP is effective in feature enhancement for all feature levels. The fact that RDP with only three feature levels significantly outperforms the baseline, further suggests that RDP is very useful to FPN.

4) Error Analyses of the Baseline and the Proposed Designs: Salient instances are usually large because large objects are more eye-attracting and are thus visually distinctive. We follow the MS-COCO benchmark to consider the instances whose areas are larger than 64^2 as large instances. In this way, we find that the ISOD dataset [7] has over 70% large salient instances. Here, we perform error analyses using all salient instances or only large instances. We view FPN [14] as the baseline and gradually add each design of us to this baseline

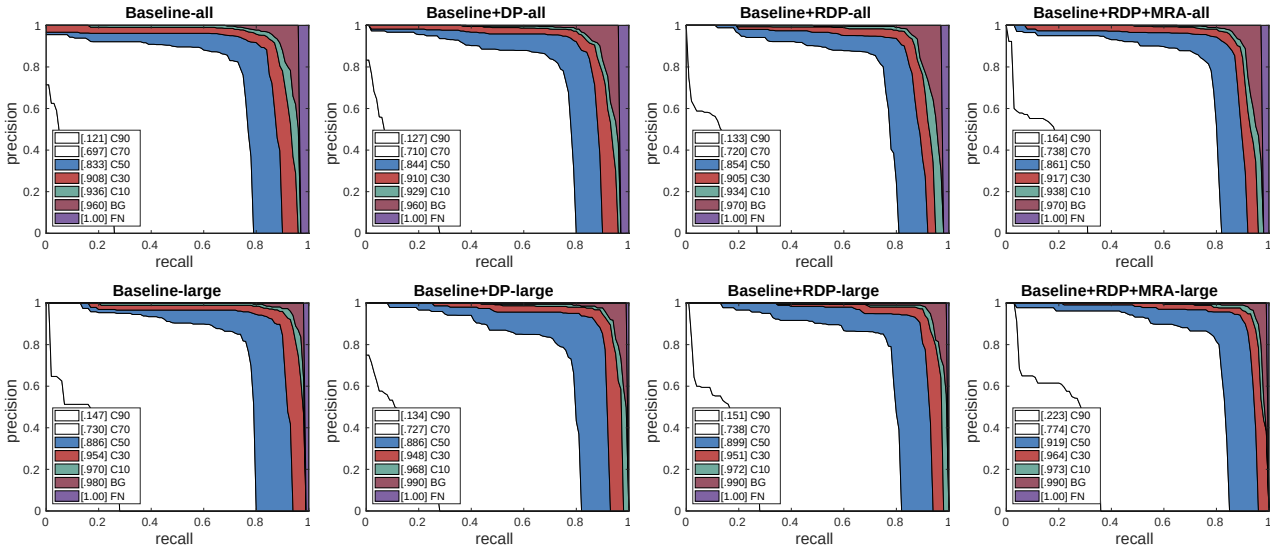


Figure 4. Error analyses for the baseline and the proposed designs on the ISOD validation set. The first row is PR curves for all salient instances, while the second row is only for large salient instances whose areas are larger than 64^2. The PR curves are drawn in different settings following [67]. C10∼C90: PR curve at IoU=$\{0.1{:}0.1{:}0.9\}$. BG: PR curve after all false positives (FP) of background are removed. FN: PR curve after all remaining errors are removed (AP = 1). Each number in the legend corresponds to the average precision for each setting. The area under each curve is drawn in different colors, corresponding to the color in the legend. Best viewed in color.

Table IV
EVALUATION ON THE ISOD VALIDATION SET FOR THE TOP-DOWN AND BOTTOM-UP DESIGNS OF RDP. THE TOP-DOWN DESIGN DIRECTLY REPLACE FPN OF THE BASELINE METHOD WITH THE TOP-DOWN STYLE OF RDP. THE BOTTOM-UP DESIGN IS THE DEFAULT VERSION OF RDP AS SHOWN IN FIG. 2.

Method	AP	AP_{50}	AP_{70}
Baseline	54.2%	83.3%	69.7%
Top-down	45.6%	76.9%	57.0%
Bottom-up	56.4%	85.4%	72.0%

Table V
COMPARISON OF DIFFERENT FEATURE PYRAMID ENHANCEMENT STRATEGIES.

Method	AP	AP_{50}	AP_{70}
Baseline	54.2%	83.3%	69.7%
+PA [58]	54.6%	84.9%	70.0%
+NAS-FPN [59]	54.3%	84.6%	69.6%
+BiFPN [60]	54.1%	84.3%	69.7%
+RDP	**56.4%**	**85.4%**	**72.0%**

to analyze the changes of detection errors. Fig. 4 illustrates the results. First, let us discuss the changes of the PR curve by adding DP to the baseline. We observe that although AP is improved for almost all IoU thresholds when using all salient instances, the performance becomes worse when only salient instances are considered, especially for large IoU thresholds (*e.g.*, IoU = 0.9). Then, we further replace DP with the regularized version of RDP. There is a significant improvement in terms of all IoU thresholds for both all and only large salient instances, demonstrating the importance of the proposed regularization for DP. At last, we analyze the effect of the multi-level *RoIAlign* (MRA) by further adding it to our system. A substantial improvement is observed, especially for large salient instances. For example, MRA brings AP improvements of 7.2%, 3.6%, and 2.0% for IoU thresholds 0.9, 0.7, and 0.5, respectively. Compared our final system (the rightmost column in Fig. 4) with the baseline (the leftmost column), the improvement is very visually significant in the PR curves in terms of all IoU thresholds.

Table VI
EVALUATION RESULTS ON THE ISOD [7] AND SOC [65] DATASETS. ALL METHODS ARE BASED ON RESNET-50 EXCEPT THE VGG-16 BASED MSRNET [7].

Method	ISOD [7]			SOC [65]		
	AP	AP_{50}	AP_{70}	AP	AP_{50}	AP_{70}
$MSRNet_{17}$ [7]	-	65.3%	52.3%	-	-	-
MS $R\text{-}CNN_{19}$ [49]	56.2%	84.2%	68.8%	35.8%	55.1%	44.2%
HTC_{19} [50]	45.4%	81.5%	55.9%	32.7%	57.6%	41.2%
$CenterMask_{20}$ [52]	54.0%	87.2%	68.7%	23.8%	39.5%	29.9%
$BlendMask_{20}$ [53]	53.6%	88.0%	67.4%	32.3%	56.2%	38.7%
$DetectoRS_{20}$ [55]	50.4%	82.7%	63.7%	24.3%	49.1%	28.4%
$SOLO_{20}$ [54]	53.5%	84.2%	65.3%	36.0%	58.1%	45.0%
$S4Net_{20}$ [12]	52.3%	86.7%	63.6%	24.0%	51.8%	27.5%
Ours	**58.6%**	**88.9%**	**73.8%**	**37.7%**	**59.4%**	**48.4%**

5) Bottom-up versus Top-down: In our method, we rebuild the feature pyramid based on the outputs of FPN. Another potential solution is to directly replace FPN with the top-down style of RDP, which would have a lower computational cost compared with our proposed design. However, the experimental results proclaim its failure. As shown in Table IV, this solution leads to substantial performance degradation, *i.e.*, over 10% lower than the default bottom-up design in terms of various metrics. Hence, we can come to the conclusion that the proposed RDP is not suitable for the top-down information flow but can only well in a bottom-up way.

6) Feature Pyramid Enhancement Strategies: Table V shows the quantitative comparison of the proposed RDP with other competitive feature pyramid enhancement strategies, *i.e.*, PA [58], NAS-FPN [59], and BiFPN [60]. We use the same baseline as in Section IV-C5. One can see that PA [58], NAS-FPN [59],

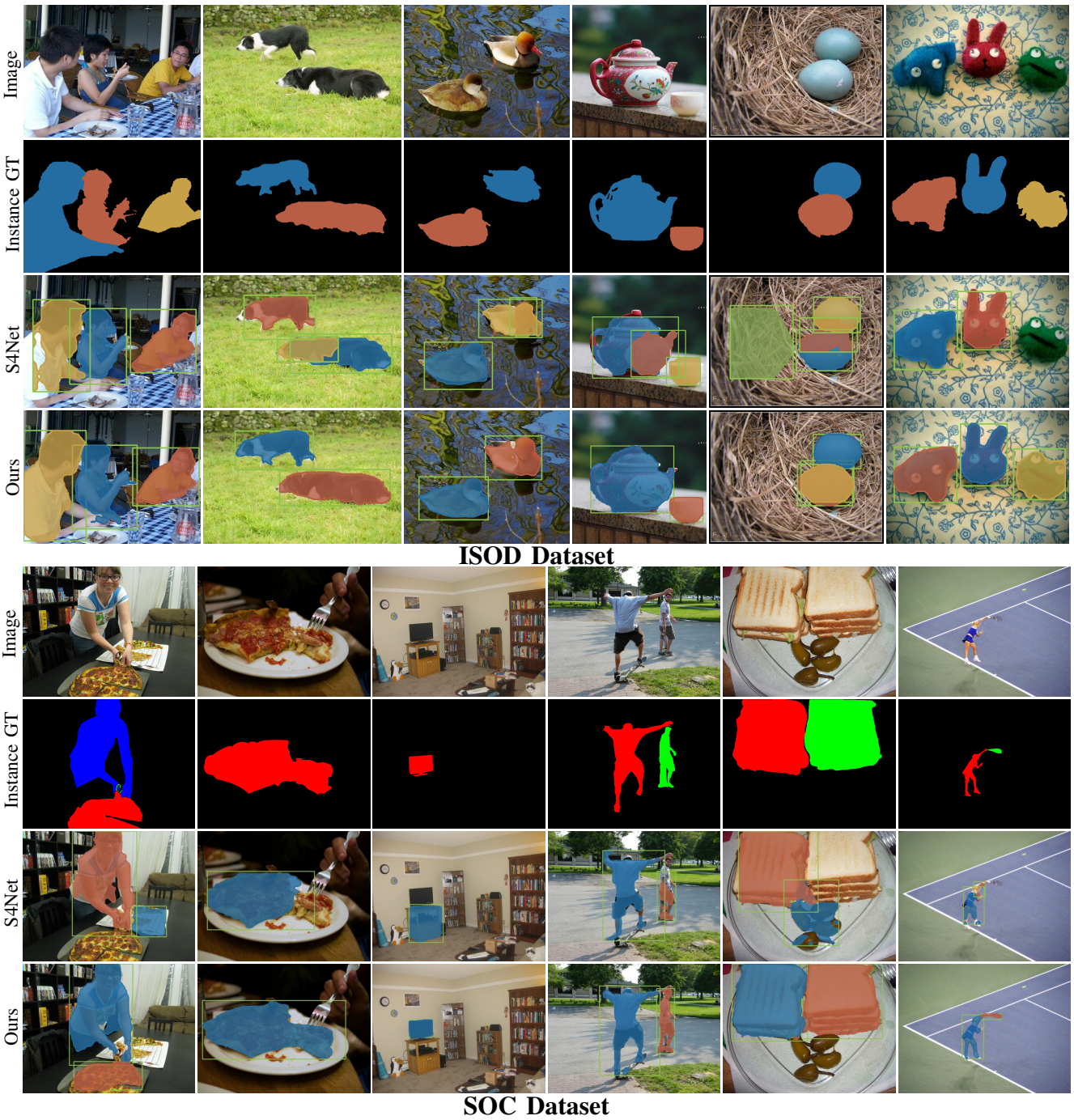


Figure 5. Qualitative comparisons between our method and S4Net [12]. The samples are from the ISOD and SOC datasets. S4Net [12] is easy to detect superfluous objects (false positives) or a part of instances. In contrast, our proposed method can detect the complete instances and have much fewer false positives.

and BiFPN [60] have a minor improvement (0.4% for PA [58], 0.1% for NAS-FPN [59]) or even no improvement (−0.1% for BiFPN [60]) over the baseline, in terms of the AP metric. In contrast, our proposed RDP outperforms the baseline by a large margin (2.2% AP improvement), demonstrating its superiority in feature pyramid enhancement.

D. Comparisons with state-of-the-art Methods

Since SIS is a relatively new problem, the previous works on this topic are very limited. Here, we compare our method with two well-known SIS methods: MSRNet [7] that is on behalf of the post-processing-based methods and S4Net [12] that is a representative work of end-to-end networks. Moreover, we compare our method with recent well-known instance segmentation methods, including Mask Scoring (MS) R-CNN [49], HTC [50], CenterMask [52], BlendMask [53], SOLO [54], and DetectoRS [55]. For a fair comparison, we train all the above methods using their official code with default settings and the ResNet-50 backbone. Hence, all methods are based on ResNet-50 [62] except the VGG-16 based MSRNet [7].

Table VII
EVALUATION OF OUR METHOD WITH DIFFERENT BACKBONE NETWORKS ON THE ISOD TEST SET [7]. OUR METHOD WITH THE MOST POWERFUL BACKBONE (*i.e.*, RESNEXT-101 [72]) CAN ACHIEVE A 4.6% IMPROVEMENT IN TERMS OF AP AND 2.7× INFERENCE TIME COMPARED WITH THAT WITH THE SIMPLEST BACKBONE (*i.e.*, RESNET-50 [62]). THE SPEED IS TESTED USING A SINGLE NVIDIA TITAN XP GPU.

Backbone	AP	AP_{50}	AP_{70}	Speed
ResNet-50 [62]	58.6%	88.9%	73.8%	45.0fps
ResNet-101 [62]	60.9%	89.7%	76.6%	34.8fps
ResNeXt-101 [72]	63.2%	90.1%	**78.1%**	16.7fps

1) ISOD Dataset: Following [7], [12], all methods are tested on the ISOD test set [7]. The quantitative results can be seen in Table VI. The proposed method achieves the best results compared with the other two popular competitors and recent strong instance segmentation methods. Specifically, the proposed method has 6.3% higher AP than S4Net [12]. In terms of AP_{70}, the proposed method is 10.2% better than S4Net [12]. Compared with recent strong instance segmentation methods, our method has a significant 2.4% improvement in terms of the AP metric. These results demonstrate the superiority of the proposed method in accurate SIS. In Table VII, we try different backbone networks for our method. One can see that powerful backbones can further significantly boost the performance, indicating the good potential and extendibility of our method.

2) SOC Dataset: The SOC dataset scenarios [65] are much more complex than those of the ISOD dataset [7], so SIS on the SOC dataset is more challenging. The quantitative comparison between our method and other recent methods on the SOC dataset is summarized in Table VI. Since other methods do not report evaluation results on this dataset, we train S4Net [12] using its official code with default settings. We leave the performance of MSRNet [7] blank due to its incomplete code. The results suggest that our method is 13.7%, 7.6%, and 20.9% better than S4Net in terms of AP, AP_{50}, and AP_{70}, respectively. Compared with recent strong instance segmentation methods, our method still has 1.7%, 1.3%, 3.4% improvement in terms of AP, AP_{50}, and AP_{70}, respectively. The above result suggests that our method can handle the cluttered background much better and our improvement for SIS is nontrivial.

E. Qualitative Comparisons

To visually compare our method with the previous state-of-the-art method of S4Net [12], we show qualitative comparisons using the ISOD [7] and SOC [65] datasets in Fig. 5. S4Net has many superfluous detection results (false positives) or only detects a part of salient instances. In contrast, our method produces consistent and high-quality results. Moreover, the boundaries of salient instances detected by S4Net are usually rough, while our method can produce salient instances with smooth boundaries. Therefore, these qualitative comparisons further validate the effectiveness of the proposed method.

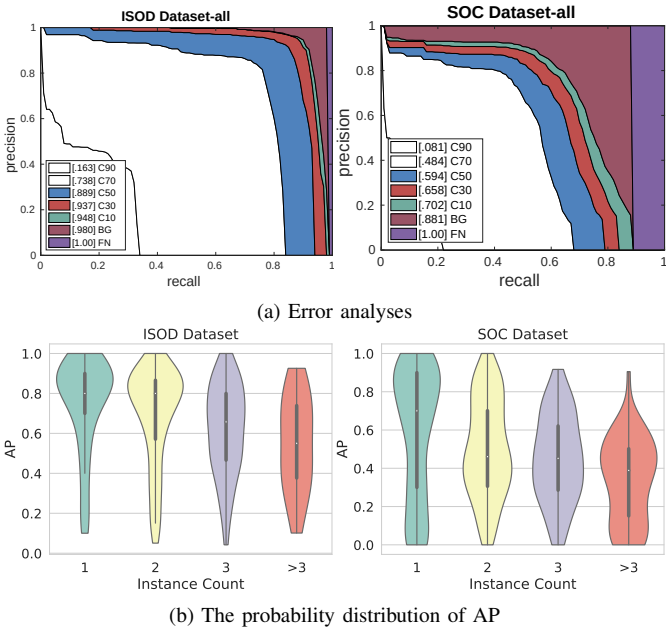


Figure 6. Statistical analyses for our method on the ISOD [7] and SOC [65] test sets.

F. Statistical Analyses

The statistical characteristics of the ISOD [7] and SOC [65] datasets are highly different, so it would be interesting to explore the differences in the performance of our method on these two datasets. Here, we conduct statistical analyses for the performance of our method on the test sets of these two datasets. We first explore the differences of PR curves between the two datasets by drawing the PR curves of our method on these two datasets, as shown in Fig. 6 (a). As the background of images in the SOC dataset is more cluttered than that in the ISOD dataset, more salient instances are not detected in the SOC dataset. In contrast, in the ISOD dataset, most salient instances can be correctly localized. Then, we explore the probability distribution of AP for different numbers of salient instances in each image. More specifically, we calculate the AP score and the number of ground-truth salient instances for each image. We then illustrate the overall probability distribution in Fig. 6 (b), where the area of each closed pattern is 1 (*i.e.*, the sum of all probabilities). $AP = 1$ for an image means that our method almost perfectly detects and segments the ground truths in this image and has no false positives. $AP = 0$ indicates that all ground truths in this image are not detected. In the ISOD dataset, each image's AP score is likely better than the medium AP score if the instance count is not more than 3 in each image, while in the SOC dataset, the same case happens only when the instance count is 1 in each image. Besides, in the ISOD dataset, our method only fails for a few images ($AP = 0$) with 1 or 2 salient instances in each image. However, in the SOC dataset, our method fails for relatively many more images. The above analyses suggest that the SOC dataset is much more difficult than the ISOD dataset due to its cluttered background and complex scenarios. Therefore, there might still be much space to strengthen the representation for future SIS research.

V. Conclusion and Future Work

In this paper, we propose a new network for salient instance segmentation (SIS). Our method's core is the regularized dense-connected pyramid (RDP), which provides each side-output with richer yet more compatible bottom-up information flows to enhance the side-output prediction. We further design a novel multi-level RoIAlign based decoder for better mask prediction. Through extensive experiments, we analyze the effect of our proposed designs and demonstrate the effectiveness of our method. With our simple designs, the proposed method achieves state-of-the-art results on popular benchmarks in terms of all evaluation metrics while keeping a real-time speed. The effectiveness and efficiency of the proposed method make it possible for many real-world applications. Moreover, this research is expected to push forward the development of feature learning and mask prediction for SIS. In the future, we plan to apply the RDP module for other vision tasks that need powerful feature pyramids. To promote future research, code and pretrained models will be released at https://github.com/yuhuan-wu/RDPNet.

Acknowledgment

This research was supported by the Major Project for New Generation of AI under Grant No. 2018AAA0100400, S&T innovation project from Chinese Ministry of Education, and NSFC (61922046).

References

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

[2] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng, "Salient object detection: A discriminative regional feature integration approach," *Int. J. Comput. Vis.*, vol. 123, no. 2, pp. 251–268, 2017.

[3] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Salient object detection with recurrent fully convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1734–1746, 2018.

[4] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2019.

[5] Y. Liu, M.-M. Cheng, X. Zhang, G.-Y. Nie, and M. Wang, "DNA: Deeply-supervised nonlinear aggregation for salient object detection," *IEEE Trans. Cyb.*, 2021.

[6] R. Fan, Q. Hou, M.-M. Cheng, G. Yu, R. R. Martin, and S.-M. Hu, "Associating inter-image salient instances for weakly supervised semantic segmentation," in *Eur. Conf. Comput. Vis.*, 2018, pp. 367–383.

[7] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2386–2395.

[8] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt *et al.*, "From captions to visual concepts and back," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1473–1482.

[9] Y. Liu, Y.-H. Wu, P.-S. Wen, Y.-J. Shi, Y. Qiu, and M.-M. Cheng, "Leveraging instance-, image-and dataset-level information for weakly supervised instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[10] Q. Hou, L. Han, and M.-M. Cheng, "Autonomous learning of semantic segmentation from internet images (in chinese)," *Sci Sin Inform*, 2021.

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

[12] R. Fan, M.-M. Cheng, Q. Hou, T.-J. Mu, J. Wang, and S.-M. Hu, "S4net: Single stage salient-instance segmentation," *Computational Visual Media*, vol. 6, no. 2, pp. 191–204, June 2020.

[13] K. He, G. Gkioxari, P. Dollr, and R. Girshick, "Mask r-cnn," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, 2020.

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

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[16] Y.-H. Wu, S.-H. Gao, J. Mei, J. Xu, D.-P. Fan, R.-G. Zhang, and M.-M. Cheng, "JCS: An explainable covid-19 diagnosis system by joint classification and segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 3113–3126, 2021.

[17] Y. Liu, Y.-H. Wu, Y. Ban, H. Wang, and M.-M. Cheng, "Rethinking computer-aided tuberculosis diagnosis," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

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

[19] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Int. Conf. Comput. Vis.*, 2013, pp. 1529–1536.

[20] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, 2014.

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

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3431–3440.

[23] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 678–686.

[24] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, p. 815, 2019.

[25] Y. Qiu, Y. Liu, H. Yang, and J. Xu, "A simple saliency detection approach via automatic top-down feature fusion," *Neurocomputing*, vol. 388, pp. 124–134, 2020.

[26] Y.-H. Wu, Y. Liu, L. Zhang, and M.-M. Cheng, "Edn: Salient object detection via extremely-downsampled network," *arXiv preprint arXiv:2012.13093*, 2020.

[27] Y.-H. Wu, Y. Liu, J. Xu, J.-W. Bian, Y. Gu, and M.-M. Cheng, "Mobilesal: Extremely efficient rgb-d salient object detection," *arXiv preprint arXiv:2012.13095*, 2020.

[28] S.-H. Gao, Y.-Q. Tan, M.-M. Cheng, C. Lu, Y. Chen, and S. Yan, "Highly efficient salient object detection with 100K parameters," in *Eur. Conf. Comput. Vis.*, 2020.

[29] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[30] Y. Liu, Y.-C. Gu, X.-Y. Zhang, W. Wang, and M.-M. Cheng, "Lightweight salient object detection via hierarchical visual perception learning," *IEEE Trans. Cyb.*, 2020.

[31] P. Zhang, W. Liu, H. Lu, and C. Shen, "Salient object detection with lossless feature reflection and weighted structural loss," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3048–3060, 2019.

[32] S. Xie and Z. Tu, "Holistically-nested edge detection," *Int. J. Comput. Vis.*, vol. 125, no. 1-3, pp. 3–18, 2017.

[33] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, and J. Tang, "Richer convolutional features for edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1939–1946, 2019.

[34] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, "Detect globally, refine locally: A novel approach to saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3127–3135.

[35] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3089–3098.

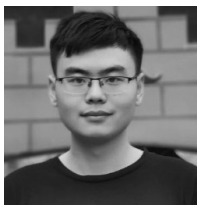
[36] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational visual media*, vol. 5, no. 2, pp. 117–150, 2019.

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

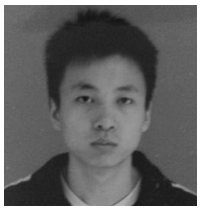
[38] Y. Wang, X. Zhao, X. Hu, Y. Li, and K. Huang, "Focal boundary guided salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2813–2824, 2019.

[39] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey." *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[40] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Int. Conf. Comput. Vis.*, 2014, pp. 580–587.
[41] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 447–456.
[42] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3992–4000.
[43] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.
[44] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 128–140, 2017.
[45] Y. Liu, S. Li, and M.-M. Cheng, "Refinedbox: Refining for fewer and high-quality object proposals," *Neurocomputing*, vol. 406, pp. 106–116, 2020.
[46] M.-M. Cheng, Y. Liu, W.-Y. Lin, Z. Zhang, P. L. Rosin, and P. H. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," *Computational Visual Media*, vol. 5, no. 1, pp. 3–20, 2019.
[47] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2359–2367.
[48] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Adv. Neural Inform. Process. Syst.*, 2015, pp. 91–99.
[49] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask Scoring R-CNN," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 6409–6418.
[50] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang *et al.*, "Hybrid task cascade for instance segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 4974–4983.
[51] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A simple and strong anchor-free object detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.
[52] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 13 906–13 915.
[53] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, "Blendmask: Top-down meets bottom-up for instance segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 8573–8581.
[54] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "Solo: Segmenting objects by locations," in *Eur. Conf. Comput. Vis.* Springer, 2020, pp. 649–665.
[55] S. Qiao, L.-C. Chen, and A. Yuille, "Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution," *arXiv preprint arXiv:2006.02334*, 2020.
[56] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, 2008, pp. 1–8.
[57] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[58] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 8759–8768.
[59] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 7036–7045.
[60] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 10 781–10 790.
[61] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.
[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.
[63] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[64] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in *ACM Int. Conf. Multimedia*. ACM, 2016, pp. 516–520.
[65] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Eur. Conf. Comput. Vis.* Springer, 2018, pp. 186–202.
[66] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (voc) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[67] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755.
[68] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Eur. Conf. Comput. Vis.* Springer, 2016, pp. 21–37.
[69] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.
[70] S.-M. Hu, D. Liang, G.-Y. Yang, G.-W. Yang, and W.-Y. Zhou, "Jittor: a novel deep learning framework with meta-operators and unified graph execution," *Science China Information Sciences*, vol. 63, no. 222103, pp. 1–21, 2020.
[71] Y. Wu and K. He, "Group normalization," in *Eur. Conf. Comput. Vis.* Springer, 2018, pp. 3–19.
[72] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 1492–1500.



Yu-Huan Wu is currently a Ph.D. candidate with College of Computer Science at Nankai University, supervised by Prof. Ming-Ming Cheng. He received his bachelor's degree from Xidian University in 2018. His research interests include computer vision and machine learning.



Yun Liu received his Ph.D. degree from Nankai University in 2020. Currently, he works as a postdoctoral scholar with Prof. Luc Van Gool in ETH Zurich. His research interests include computer vision and machine learning. He has published over 10 papers in IEEE TPAMI, IJCV, IEEE CVPR, IEEE ICCV, *etc*. He has received the National Scholarship in 2017, 2019, and 2020 three times.

Le Zhang received his M.Sc and Ph.D.degree form Nanyang Technological University (NTU) in 2012 and 2016, respectively. Currently, he is a scientist at Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. He served as TPC member in several conferences such as AAAI, IJCAI. He has served as a Guest Editor for Pattern Recognition and Neurocomputing; His current research interests include deep learning and computer vision.

Wang Gao received his master degree from The Third Research Institute of China Aerospace Science and Industry Corporation in 2017. He is currently with the Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, Beijing, China. His research interests include computer vision, scene matching and visual navigation.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012. Then he did two years research fellow with Prof. Philip Torr in Oxford. He is now a professor at Nankai University, leading the Media Computing Lab. His research interests include computer graphics, computer vision, and image processing. He received research awards, including ACM China Rising Star Award, IBM Global SUR Award, and CCF-Intel Young Faculty Researcher Program. He is on the editorial boards of IEEE TIP.