Weakly Supervised Learning with Region and Box-level Annotations for Salient Instance Segmentation

Jialun Pei, He Tang and Chuanbo Chen

Abstract—Salient instance segmentation is a new challenging task that received widespread attention in saliency detection area. Due to the limited scale of the existing dataset and the high mask annotations cost, it is difficult to train a salient instance neural network completely. In this paper, we appeal to train a salient instance segmentation framework by a weakly supervised source without resorting to laborious labeling. We present a cyclic global context salient instance segmentation network (CGCNet), which is supervised by the combination of the binary salient regions and bounding boxes from the existing saliency detection datasets. For a precise pixel-level location, a global feature refining layer is introduced that dilates the context features of each salient instance to the global context in the image. Meanwhile, a labeling updating scheme is embedded in the proposed framework to online update the weak annotations for next iteration. Experiment results demonstrate that the proposed end-to-end network trained by weakly supervised annotations can be competitive to the existing fully supervised salient instance segmentation methods. Without bells and whistles, our proposed method achieves a mask AP of 57.13%, which outperforms the best fully supervised methods and establishes new states of the art for weakly supervised salient instance segmentation.

Index Terms—Weakly supervision, saliency detection, instance segmentation, deep learning.

I. INTRODUCTION

SALINET object detection (SOD) is known as a classic research field for highlighting the most sensitive and informative regions of an image [1]–[3]. Originating from cognitive and psychology research communities, saliency detection is applied to various areas, such as image captioning [4], video summarization [5] and content-aware image editing [6]. Albeit salient object detection task provides the salient proposal labels compared to the background, they do not explore instance-level salient information. The next generation of salient object detection methods need to showcase more detailed parsing and identify individual instances in salient regions [7]. In addition, instance-level salient information is more consistent with human perception and offers better image understanding [8]. In this paper, we focus on the new challenging task salient instance segmentation (SIS) and segment salient instances accurately based on the salient object detection.

Saliency detection has gained significant progress owing to the rapid development of deep convolutional neural networks (CNNs) [9]–[11]. Driven by the strong capability of multi-level feature extraction, CNN models are widely used in object detection and instance segmentation [12], [13], especially focusing on estimating the bounding boxes of salient instances [14]. Based on the above contribution and feasibility, salient instance segmentation can produce pixel-wise salient instances [15]. Unlike instance segmentation, salient instance segmentation only predicts salient instances based on the salient regions. Besides, segmenting salient instances is class-agnostic compared to the class-specific instance segmentation task. Different from salient object detection, salient instance segmentation fosters a more detailed information by labeling each instance with a precise pixel-wise mask.

However, the performance of CNNs is mostly dependent on requiring the pixel-level fully-supervised train data [16], [17]. Up to now, pixel-wise annotated SIS dataset is seriously inadequate and the amount of pixel-wise ground-truth images is not enough. The quality and quantity of pixel-level annotations is the bottleneck because the labeling task is strenuous and time-consuming. To alleviate the influence of lacking fully-supervised data, weakly supervised learning can be seen as the alternative training method attracting more attention. It not only avoids user-intensive labeling, but also allows the neural network to get sufficient training samples.

Inspired by this consideration, in this paper, we aim to combine the salient instances of bounding boxes and binary salient region maps for weakly supervised learning of salient instance segmentation. Not only the cost of bounding box annotation is low, but also contain location information for each salient instance. Besides, the ground truth of salient region map is a ready-made source generated from a large collection of SOD datasets and it provides salient region information for salient instance segmentation. As showed in Fig. 1, we re-labeled the instance-level bounding boxes in an image to determine the location and number of salient instances instead of the existing bounding boxes of salient regions which have labeled in the DUT-OMRON dataset produced by Yang et al. [18]. We use the proposals of salient instances and saliency maps to assign salient regions to each bounding box of salient instance. If a closed salient region happens to correspond to a saliency instance, the salient instance label is viewed as an intact label, which can refer to the salient instance in the lower left corner of the sample image in Fig. 1a. In order to ensure one instance correspond to one bounding box and the consistency of salient instances and regions, we also exploit some priors to limit the different object regions trapped into same box as much as possible. In this case, the network can achieve more training samples.
with the lowest labeling cost so as to relieve the shortage of segmentation supervision. We will elaborate the generation steps of weak annotations in Section III.C.

For segmenting salient instances, we design a cyclic global context salient instance segmentation neural network (CGCNet) supervised by the above homemade weak labels. Fig. 2 shows the overview of our CGCNet. The proposed approach is an end-to-end two-stage framework, which first detects salient proposals and then predict the pixel-level salient instance masks. When extracting features for mask prediction, the performance of convolutional layer depends heavily on global context. Considering better feature representation, we expend the scope of feature extraction from the local proposal to the global features [19]. Inspired by enter-surround contrast derived from saliency detection mechanism [20]–[22], a global feature refining module (GFR) is designed to make full use of background features and suppress disturbance from other salient instance features. Different from the ROIAlign layer that limits the receptive field in Mask R-CNN [23], the proposed GFR module is sensitive to global contrast so as to capture more detailed edge information. Moreover, the CGCNet is designed to iteratively update the weak annotations by the forward prediction masks combining with a conditional random field (CRF) [24], which is benefit to refine the weak annotations sequentially. The input training sample and the corresponding result is shown in Fig. 1. We evaluate the results on the test set of Dataset1K [7] and show that our method performs significantly on salient instance segmentation.

In summary, the main contributions of this paper are as follows:

- We present an end-to-end cyclic global context neural network (CGCNet) for salient instance segmentation. The framework is class-agnostic and achieve accurate salient instance results. To the best of our knowledge, we are among the first to train a weakly supervised salient instance segmentation network.
- The proposed model is weakly supervised by our homemade weak labels, which are made by the combination of the region-level bounding boxes and salient regions. In addition, an update scheme can optimize the weak ground-truth continuously to improve the accuracy.
- We design a global feature refining layer (GFR) layer which extend the receptive field of each instance to the global context and suppress the features of other instances simultaneously.

The remainder of this paper is organized as follows. Section II presents the related works. Section III describes the architecture and the details of the proposed framework. Section IV discusses the experimental settings and comparisons with the state-of-the-art methods. Finally, Section V concludes the paper.

II. RELATED WORK

A. Salient Object Detection

Thanks to the fast development of deep learning techniques, salient object detection has gone through a transformation from traditional machine learning to deep CNNs [25]. Driven by the multi-level features extracted from convolution network, the saliency maps are predicted accurately and efficiently. Fortunately, rich pixel-level salient datasets can be poured into various CNN models to detect salient regions [26]–[28]. Li et al. [29] proposed a multi-scale deep contrast network to overcome the limitations of overlap and redundancy. Hou et al. [30] designed short connections to the skip-layer structures based on the VGGNet for better supervision. Qin et al. [31] produce a predict-refine SOD network which is composed of a densely supervised Encoder-Decoder network and a residual refinement module. Although these SOD methods achieved outstanding performance, the saliency map is considered as the region-level binary mask which may not accomplish instance-level salient object segmentation.

B. Salient Instance Segmentation

On the basis of SOD, salient instance segmentation pushes the problem into an instance-level phase. Unlike instance segmentation [32]–[34], salient instance is category-independent and it is limited into the salient regions. Therefore, the network and data set of instance segmentation are incompatible with segmenting salient instances. SIS task should address two crucial problem: (1) Existence: determine if the salient region has one salient instance and enumerate the number of salient instances. (2) Location: if a salient instance exists, we should determine where each instance is located. Zhang et al. [14] generated salient region-level proposals by CNNs and optimized the bounding boxes based on the Maximum a Posteriori principle. The method is the first raise saliency detection from the region level to the instance level. Subsequently, Li et al. [7] formally proposed the instance-level salient object detection task. The author promoted the results of SIS from proposals...
Fig. 2: An overview of the proposed framework. The detail of the GFR module is shown in the upper right corner. The weak annotations updating criteria is illustrated in Section III.D. At the training time, the salient instance result return to update the weakly supervised annotation in next iteration.

to pixel-level and hand-crafted the first SIS dataset containing 1,000 samples. Recently, Fan et al. [15] proposed an end-to-end single-shot salient instance segmentation framework to segment salient instances. The proposed ROIMasking layer allow more detailed information to be detected accurately, and meanwhile remain the context information around the regions of interest [15]. As a new challenging task, however, the lacking of fully-supervised label is the main problem to limit the performance of deep learning models for segmenting salient instances. To avoid making the high cost of pixel-level annotations, we take advantage of weakly supervised learning for training our model.

C. Weakly Supervised Learning

Most of neural networks require fully supervision in the form of handcrafted pixel-level masks which limits their application on large-scale datasets with weaker forms of labeling [35]. To reduce the cost of hand-labeling, weakly supervised learning has attracted a great deal of attention in recent years [36–38]. Many weakly supervised principles have been introduced in computer vision area, including object detection, instance segmentation and saliency detection [39], [40]. Weakly supervision reveals that the deep learning network purposed for one supervision source can resort to another source or incomplete labels. Li et al. [41] utilized a coarse activation map from the classification network and saliency maps generated from unsupervised methods as pixel-level annotation to detect salient object. Wang et al. [42] developed a Foreground Inference Network (FIN) for saliency detection, which was supervised by using image-level tags only. Moreover, Zeng et al. [43] incorporated with diverse supervision sources to train saliency detection models. They designed three networks that learn from category labels, captions and noisy labels, respectively. Inspired by the above contributions, we combine with the existing binary salient regions and bounding boxes to train a more robust neural network.

III. THE CGCNET ARCHITECTURE

A. Motivation

The motivation of the proposed method is handled with segmenting class-agnostic salient instances under lacking fully-supervised annotations. We tend to utilize as many training samples as possible with the lowest labeling cost. Therefore, in this paper, the weak annotations are composed of a combination of salient proposals and binary salient regions. On one hand, the salient proposals are easily available and provide positional information. On the other hand, binary salient regions can provide approximate salient area information for salient instances and they can easily obtain from a large number of existing saliency detection datasets. For salient instance segmentation, it belongs to the scope of saliency detection, which is category-independent even though it is promoted to the instance-level task. So, we think it is insignificant to train the salient instance network by the image-level annotations.

For training by the weak labels, we design a cyclic global context neural network (CGCNet) to predict salient instances and update the weak labels recurrently. In this section, we elaborate on the proposed weakly supervised framework for SIS.

B. Overall Framework

As shown in Fig. 2 the framework of our proposed CGCNet is clear and understandable in the whole process. The end-to-end framework consists of three main components. Firstly, The RPN head is seen as a salient proposal detector to capture
the location and number of salient instances. It produces the bounding boxes of salient instance for the segmentation branch which produces the pixel-level salient instances. It’s worth noting that the GFR module provides the global feature representation to predict salient masks. Not only that, the resulting salient instances update the weak ground-truth added with the fully connected CRF operation for the next iteration.

Like most recent instance segmentation works, we combine ResNet-101 [44] with FPN [45] as our pre-trained backbone. According to the order of down sampling in ResNet-101, we extract the 4-th stage feature map followed by a 1 x 1 convolutional layer with the lateral connections in multi-level FPN prediction [23]. Followed by FPN, we mask use of five levels of feature maps to detect different sizes of objects on different levels so as to maximize the gains in accuracy. It’s worth noting that the feature maps produced by the backbone are extracted from the entire input image. Both salient proposal detector and salient instance segmentation branch will be feed with the 256 channel feature maps.

Similar to Faster R-CNN [46], the RPN head is merged into CGCNet for predicting the bounding boxes of each instance in one image. Considering to the character of category-independent, each ROI feature is assigned to two classes, denoted as $B_c \in \{0, 1\}$. The two classifications correspond to the background and the salient object in foreground. Salient RPN works on the input features and predicts a pile of salient proposals. Followed by ROIAlign [23] and two 1024-D Fully Connected layer (FC), the resulting coordinates of salient proposals are generated attached with a confidence score of saliency degree. Then, Non-Maximum Suppression (NMS) [47] is set to suppress the negative proposals which the saliency score behind the threshold 0.7 for refining the bounding box of each instance.

The output salient proposals relabel on the feature maps produced by the backbone as input to our GFR module. In this process, the GFR module extends the ROI feature to the global feature. In addition, this layer retains the feature of the current instance while suppressing the feature of other ROI features. The feature processed by the GFR module are injected into a pixel-to-pixel fully convolutional block. The Fully convolutional fashion can preserve the spatial consistency of each pixel involved in corresponding salient instance. More importantly, Taking the resulting salient instances predicted by the SIS branch, the updating scheme is constructed to replace the weak ground-truth recurrently in training time. In following subsection, we will describe the SIS branch and the GFR module in detail.

C. Weak Supervision Sources

In order to handle with lacking datasets for SIS task, we implement weakly supervised learning for training our network. Considering the characteristics of salient instances, we need embrace both the pixel-level salient region like saliency detection and the number of salient instances. Inspired by salient object detection and instance segmentation task, our weak labels are composed by salient regions and the bounding boxes of salient regions. We use about 4,500 images from the training set of DUT-OMRON SOD dataset because it has a sufficient number of ground-truth and the number of instances contained in the image is distributed uniformly. Besides, the dataset also has the bounding boxes labels. But the bounding boxes labeled according to the salient object regions rather than based on the salient instances. Consequently, we manually annotated salient instance bounding boxes according to the location and number of salient instances in one image. In order to ensure consistency and fairness, we collected five annotations per instance from distinct workers and then take the average value [48]. To ensure the quality of weak labels, we excluded the image which have no salient instances and some ambiguous cases in the process of marking. Each salient bounding box is annotated by a 4-D vector composed of the normalized coordinates of its upper-left and bottom-right corners [14]. In spite of combining salient regions and bounding boxes, the weak labels still have some general issues. Firstly, salient regions from different bounding boxes have shared patches. Secondly, some small instances are enclosed into the bounding boxes of larger instances. To reduce the negative influence of these obstacles, we provide two priors to deal with ambiguous samples. On one hand, we restrict that each bounding box can contain only one enclosed salient region. On the other hand, if there are multiple closed areas in one bounding box, we only keep the maximal area as its regression target. Given a binary salient map $S$, the manually marked window is set as $W_i(i = 1, 2, ..., \infty)$. In addition, we set the patches discarded by priors in each window as $\theta$. The final weak label $I$ is defined by:

$$ I = \sum_{i=1}^{n} m[W_i[S(x, y); \theta(x, y)], \; i = 1, 2, ..., n}$$

(1)

where $(x, y)$ presents salient region pixels in image $S$. The final example can refer to Fig. 1.

D. The Salient Instance Segmentation Branch

The salient instance segmentation branch aims to segment each salient instance in virtue of the global cues. Achieving the features of regions from the RPN head, we can determine the location and number of salient instances. However, features of each region just contain local spatial information, which is insufficient to segment explicit pixel-level labels. This barrier drives us to explore the broader feature for segmenting. Inspired by center-surround contrast based on SOD task, we seek to expend the ROI feature to the global feature map. Resorting to maintain as many features as possible and ensure the resolution of instances, we take advantage of the global features extracted from the backbone instead of the ROI feature. Meanwhile, each feature map produced from GFR module only contain the feature of background and current salient instance proposal while suppressing other salient instance proposals.

The GFR module. As mentioned earlier, the goal of the proposed global feature refining module (GFR) is to obtain global context information and limit the disturbance of other instance features. Compared with ROIAlign in instance segmentation area, they only pay close attention to the ROI feature and resize the original resolution of ROI [23]. In S4Net
Fig. 3: Visualization of the GFR module in segmentation branch and comparison of our local feature refining module (LFR module) and Mask R-CNN [23]. Salient proposals are produced by the RPN head and the feature map is extracted by the backbone from CGCNet.

[15], the ROI Masking extend the receptive field and use of the information around the ROI contrasting ROI features. Differ from the ROI Masking, our GFR module expand each ROI directly to the global feature map and maximize the center-surround contrast for segmenting salient instance.

The internal process in the GFR module is shown in the top right corner of Fig. 2. Given the feature maps produced from FPN, the GFR module first map the coordinates of all top right corner of Fig. 2. For better verification, we visualized the process of GFR in Section IV.C. The experiment results demonstrate that the GFR module outperforms the other two modules for salient instance segmentation task, which is discussed in detail in Section IV.C.

By means of GFR module, each goal of instance not only contains the features inside the proposal but also take advantage of the global context information so as to make the instance region prominent. The next FCN-based network can better use the contrast of foreground and background features to segment salient instances. For each output feature map from GFR layer, SIS branch stack four consecutive convolutional layers followed on dilated convolutional layer with stride 2 and RELU function [49]. All the convolutional layers have a kernel size 3×3 and stride 1. Experimental results demonstrated that it is efficient and lightweight to segment each salient instance.

Weak Annotations Updating Scheme. Considering the initial training samples are weak annotations, we leverage an updating scheme to optimize weak annotations continuously. The fundamental flaw of our weak labels is that boundary information of each instance is not have sufficiently detailed and different instances in one image have overlap and occlusion. If just training on the original samples, the resulting salient instances would contain some small redundant patches which belong to background or other instances. To further improve the performance of CGCNet, we insert the fully connected Conditional Random Field (CRF) [24] after the salient instance maps in the SIS branch because the CRF operation has significant progress on refining the edge of objects. The fully connected CRF model employs the following energy function:

\[
E(M) = - \sum_i \log P(m_i) + \sum_{i,j} \varphi_{\theta}(m_i, m_j) \tag{3}
\]

where \(M\) presents a binary mask assignment for all pixels, and \(P(m_i)\) is the label assignment probability at pixel \(i\) belonging...
to the salient instance. For each binary salient instance mask, the pairwise potential $\varphi_p(m_i, m_j)$ for two labels $m_i$ and $m_j$ is defined by:

$$
\varphi_p(m_i, m_j) = \omega_1 \exp\left(-\frac{|p_i - p_j|^2}{2\sigma^2}\right) - \frac{|I_i - I_j|^2}{2\theta^2} + w_2 \exp\left(-\frac{|p_i - p_j|^2}{2\theta^2}\right) \right)^2
$$

(4)

where the first kernel depends on pixel positions ($p$) and pixel intensities ($I$), and encourages nearby pixels with similar features to take similar salient instance labels [29]. The second kernel quantifies the smoothness kernel which only depends on pixel positions for removing small isolated regions [50]. $\omega_1$ and $\omega_2$ indicate the weighted values to balance the two parts. The hyper parameters $\theta_\alpha$, $\theta_\beta$ and $\theta_\gamma$ control the degree of the Gaussian kernels. In this paper, we adopt the publicly available implementation of [24] to optimize these parameters. Specifically, we cross-validate the hyperparameters $\omega_\alpha$, $\omega_\beta$, $\theta_\alpha$, $\theta_\beta$ and $\theta_\gamma$ for the best performance of CRF. The coarse-to-fine scheme is applied on the subset of validation set (about 100 images) in DUT-ORMON dataset. The default value of $\omega_\beta$ and $\theta_\alpha$ are set to 3 and 1, and the initial search range of the parameters are $\omega_\alpha \in [1:10]$, $\theta_\alpha \in [50:5:100]$ and $\theta_\beta \in [5:1:15]$. These parameters are fixed the number of mean field iterations to 10 for achieving the best values. Finally, the values of $\omega_1$, $\omega_2$, $\theta_\alpha$, $\theta_\beta$ and $\theta_\gamma$ are set to 4, 3, 70, 13, 1, respectively in our experiments.

We denote the salient instance map as $R$ and the map processed by CRF as $R_f$. The weak annotation is labeled as $C$. According to Algorithm 1, we proposed a strategy based on the KL-Divergence [31] to selectively take the place of $C$ for the next iteration. KL-Divergence is defined as dissimilarity metric and a lower value indicates a better approximation between the predicting salient instance maps and the ground-truth. Due to ground-truth of CCGNet is noisy, however, the updating map should have more dissimilarity patches with weak annotation as well as the larger value of KL-Divergence between themselves. Our strategy compares result image $R$ and $R_f$ to the weak annotation $C$, which is designed as followed:

$$
K_1(R, C) = \frac{1}{H \times W} \sum_{i=1}^{H \times W} C_i \log\left(\frac{C_i}{R_i + \sigma}\right) + \sigma
$$

(5)

$$
K_2(R_f, C) = \frac{1}{H \times W} \sum_{i=1}^{H \times W} C_i \log\left(\frac{C_i}{R_f + \sigma}\right) + \sigma
$$

(6)

where $K_1$ and $K_2$ denote the mean KL-Divergence value of $R$ and $R_f$ to $C$, respectively, the index of $i$ is set as the $i$-th pixel and $\sigma$ is a regularization constant. In Algorithm 1, $C_n$ represents the ground-truth to be used for the next iteration. It is worth noting that we set $\varphi$ as the threshold to determine whether to update with the existing weak annotation $C$ and the value of $\varphi$ is set to 0.05. The strategy can eliminate unnecessary replacements, while limiting the impact of excessive erosion of the CRF on the resulting image. By using the updating scheme to the weakly supervised learning, the network achieved more accurate results at the training time.

**Loss Function**. The number of 4,500 training set of weak labels are used as training data. Having trained both salient proposal branch and salient instance segmentation branch simultaneously, we use manually marked bounding boxes to supervise the RPN head and the pixel-level weak labels to train SIS branch. Therefore, the loss function of CCGNet is defined as a two-stage fashion on each weak annotation:

$$
L = L_{bb} + L_{seg} + L_{upd}
$$

(7)

Here, the $L_{bb}$ function includes a classification loss which is log loss over two classes including saliency or background and a bounding box loss which is similar with $L_{loc}$ in Fast R-CNN [52]. The SIS branch loss $L_{seg}$ is defined by the cross-entropy loss, which is followed by:

$$
L_{seg} = -\frac{1}{N} \sum_{i=1}^{N} (g_i \log p_i + (1 - g_i) \log(1 - p_i))
$$

(8)

where $p_i$ denotes the probability of pixel $i$ belonging to class $c = 0, 1$, and $g_i$ indicates the ground truth label for pixel $i$. Inspired by the updating criterion from Eq. (5) and (Eq. (6), the loss function $L_{upd}$ for updating SIS branch for pixel-level salient instance prediction is:

$$
L_{upd} = K_2(R_f, C) - K_1(R, C)
$$

(9)

More importantly, the CCGNet update the RPN detector and the SIS branch with the parameters of backbone being unchanged. The entire procedure is repeated iteratively for training.

**IV. EXPERIMENTAL RESULTS**

In this section, we will elaborate on the results of proposed CCGNet framework for salient instance segmentation task in detail. Firstly, we perform detailed ablation experiments on various components of our approach. Secondly, we use different metrics to compare with the experimental results of other state-of-the-art methods. Since the proposed method is the first to accomplish SIS task by weakly supervised learning, we will try best to show the efficiency and superiority of our method as comprehensively as possible.

**A. Implementation Details**

As described in the section above, the end-to-end CCGNet is trained by our weak labels which choose 4,500 salient object images from DUT-OMRON dataset [18] without ambiguous samples. During training, the manually marked salient bounding boxes from weak labels is used as ground-truth for supervising the salient proposal detector while combining with pixel-level annotations to train SIS branch. Meanwhile, we labeled 500 images as same as training data for validation. For training salient proposals, the bounding boxes is considered.

**Algorithm 1 Weak annotations updating**

**Input:** Weak annotation $C$, salient instance map $R$ and salient instance map with CRF $R_f$.

**Ensure:** The updated weak annotation $C_n$

1: if $K_2(R_f, C) - K_1(R, C) \geq \varphi$
2: then $C_n = C$
3: else $C_n = R_f$
4: end if
as a positive sample if the IOU is more than 0.7 or a negative sample below 0.3. In addition, NMS threshold used on proposal detector is set to 0.7. In order to speed up training convergence, we initialize the CGCNet with a pre-trained model over the ImageNet dataset \cite{53} from Mask R-CNN \cite{23}.

At inference time, we only use 300 images from testing set in the dataset proposed in \cite{7} due to short of datasets for the new challenging task. In GFR module, we input the number of top 80 scoring proposals from proposal prediction branch after applying NMS. Besides, the SIS branch directly output the resulting images without the updating scheme.

Our proposed framework is implemented in PyTorch framework on 2 NVIDIA GeForce GTX 1080Ti GPUs with 22 GB of memory. The CGCNet is fine-tuned by flipping the training sets horizontally at a probability of 0.5. In our experiments, we train our network with a learning rate of 0.0025 which is decreased by 10 at the 8k iteration. The training process totally iterates 16k times by using the batch size of 4. The weight decay is empirically set to 0.0001 and the momentum is 0.9.

### B. Evaluation Metrics

For a novel and challenging task, salient instance segmentation has few evaluation metrics to measure its performance quantitatively. Unlike saliency detection, salient instance detection also distinguishes different pixel-level instances based on salient region. Compared to instance segmentation, the task pays attention to salient visual area without category. Consequently, we adopt the AP metric to calculate the average of maximum precision value at IoU scores of 0.5 and 0.7 instead of MAP metric \cite{56}. The precision value of one image is computed by the predicted number of salient instances (IoU > 0.5 or 0.7) divided by the real number of salient instances in the image. So, the \( AP' \) metric is defined by the summation of precision value divided by the number of all images in testing set, which is computed as:

\[
AP' = \frac{1}{N} \sum_{j} \frac{1}{n} \sum_{i} \text{precision}, \quad \text{IoU}(i) \geq \alpha \tag{10}
\]

\[
\text{precision} = \begin{cases} 
1, & \text{if IoU}(i) \geq \alpha \\
0, & \text{if IoU}(i) < \alpha
\end{cases} \tag{11}
\]

where \( \alpha \) is the threshold of IoU. \( N \) is the number of instances in one image and \( n \) is the total number of instances in dataset. Moreover, the \( AP \) metric is used to measure the effectiveness of salient instance segmentation according to \( AP' \) metric. The metric average the \( AP' \) metric with the threshold of IoU from 0.5 to 0.95 by step 0.05, which is calculated by:

\[
AP = \frac{1}{10} \sum_{\alpha} AP'|\alpha, \quad \alpha = 0.5, 0.55, ..., 0.95 \tag{12}
\]

Comparing with the \( AP' \) metric, the \( AP \) value can measure the overall performance of salient instance segmentation methods as a whole. In this section, the experimental results are evaluated mainly on the basis of the above two indicators.

### C. Ablation Studies

In this section, we analyze the effectiveness of the proposed CGCNet on DUT-OMRON validation set \cite{18}. To this end, the ablation studies contains four parts: performance of four different backbones, performance of GFR module and three related structures, hyper-parameter in updating scheme and contribution of each component of the architecture.

**Backbone:** To ensure fairness and the influence of the different backbones on the experimental results, we verify various mainstream backbones working on CGCNet which stay same settings. Table \[I\] shows the effectiveness of these base models working on the framework. It demonstrates that the backbone of combining ResNeXt-101 achieves the best performance whether \( AP \) or \( AP' \) metric \cite{56}. The widely used ResNet-101 has also achieved good results slightly behind ResNeXt-101. Due to insufficient network depth, VGGNet obtained relatively low accuracy, but is slightly faster than ResNet \cite{55}.

**The GFR Module:** The proposed GFR module is the core layer in SIS branch for refining features. In this paper, we have tried to test the feature refining layer containing local and global cues, respectively. Table \[II\] lists the performance of the LFR module and GFR module. Meanwhile, we also compare similar methods embedded in the segmentation branch based on CGCNet, including ROIAAlign in Mask R-CNN \cite{23} and ROIAlign in S4Net \cite{15}. As shown in Table \[II\], the experimental results based on GFR module outperforms other modules. ROIAAlign is limited to focusing only on the ROI features. Albeit LFR module expanded the scale of features around ROI, it still slightly behind the ROIAlign by reason of its ternary masking. It indicates that treatment of features by these modules play an important role in segmenting salient instances, and we finally choose the GFR module embedded in our framework.

**Hyper-parameter in updating scheme:** The threshold \( \varphi \) of updating scheme is important for the quality of weakly supervised annotations to train our framework. In our experiment, we find the appropriate threshold to ensure the efficiency at the training time. According to the formulation of KL-Divergence \cite{24}, we empirically provide several default values for determining its influence in this experiment, which is shown in Table \[III\]. The influence of different value of \( \varphi \) is
Fig. 4: Qualitative analysis of experimental results by the proposed method and S4Net [15]. The salient instance results are category-independent, each color represents only one instance in one image.

relatively average. The best result is obtained when the value of $\varphi$ was set to 0.05, it can balance the optimal quantity and quality of updates.

The component in CGCNet: It is essential to evaluate the contribution of each component to the overall framework. In this experiment, we add each innovative module to the framework one by one to discover their contributions under the same settings. These parts of CGCNet include the prior criteria (Standardized weak labels), the updating scheme and the GFR module. As shown in Table. IV the various parts of our framework have various degrees of contribution for segmenting salient instances. Particularly, the updating scheme has more contributions which improved the $AP$ metric about 2 percent compared to without it, which can be attributed to the insertion of CRF and the revision of the weak annotations at the training time. With the help of the prior criteria, the
TABLE III: The threshold $\phi$ of updating scheme performance of CGCNet. The highest scores in each row are labeled in bold.

| $\phi$ | 0.01 | 0.05 | 0.1 | 0.15 | 0.2 |
| --- | --- | --- | --- | --- | --- |
| AP$^r$ | 85.98 | 86.14 | 84.95 | 84.76 | 84.23 |
| AP | 71.43 | 71.82 | 71.26 | 70.78 | 70.29 |

TABLE IV: Ablation analysis of the influence of various components from our model on SIS task. PC, GFR and US means the prior criteria, the GFR module and the updating scheme, respectively. The experiment is evaluated on DUT-OMRON validation set.

| Models | AP | AP$^r$ | AP$^r$ |
| --- | --- | --- | --- |
| The basic model | 53.93 | 83.44 | 66.86 |
| The basic model + PC | 54.67 | 85.81 | 68.43 |
| The basic model + PC + GFR | 55.84 | 85.15 | 70.54 |
| The basic model + PC + GFR + US | 57.79 | 86.14 | 71.82 |

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

So far, there are two existing methods related with salient instance segmentation task: MSRNet [7] and S4Net [15]. Unlike these previous works, we are the first to make use of weakly supervised network for the new challenging task. In this section, all methods are evaluated on the test set of Dataset1K [7] and SOC dataset [57], respectively. For a relatively fair comparison, we compare the existing salient instance segmentation methods qualitatively and quantitatively on the only two datasets.

Evaluation on the Dataset1K: The Dataset1K [7] is the first salient instance dataset, which contains 500 images for training, 200 images for validation and 300 images for testing. Note that the MSRNet and S4Net are trained on training set of dataset1K, and the proposed method are trained on the weak annotations of DUT-OMRON dataset. Except that, the proposed CGCNet use ResNet-101 as backbone to stay the same with S4Net. Other settings also maintain relative consistency and fairness in this experiment. Table. V lists the value of AP and the results of the AP$^r$ of the metrics with IoU scores of 0.5 to 0.9 generated by different existing methods. Due to the related code of [7] is not available, we cannot get its whole results. In the case of training on the weak labels, the value of AP and AP$^r$ achieved by CGCNet surpass the S4Net, while the AP$^r$ of 0.5 metric is slightly lower than S4Net. As a weakly supervised method, our method improves the AP value of salient instances to the highest 57.13%. Since MSRNet is not an end-to-end framework and was proposed earlier, it is inferior to S4Net and the proposed method in performance.

We also qualitatively analyzed the experimental results produced by CGCNet and S4Net. Fig. 4 displays some sample images from the testing set in [7]. It can be seen that our method produces high quality results which is very close to the ground-truth. The first two input images contain two instances, which have similar internal features and relatively simple backgrounds. Our method can easily segment salient instances from the background. The middle images in Fig. 4 have multiple instances and each instance is close together. The proposed model can still predict the number of instances accurately and segments them effectively. The last two samples have chaotic backgrounds, and the internal features of salient instances are also very messy. In this complex case, the CGCNet can well distinguish obstructed instances and segment them satisfactorily. In comparison, the S4Net cannot accurately determine the number of salient instances in some cases. The antepenult sample demonstrated that the S4Net is not more sensitive to smaller salient instances than our framework. In addition, our method is better than S4Net in smoothing the edge of salient instances. It indicates that the lack of fully supervised data limits the effectiveness of S4Net. By and large, the proposed framework has high accuracy and generalization for salient instance segmentation task.

Evaluation on the SOC: Recently, Fan et al. [57] in-

Fig. 5: The attributes-based performance of the CGCNet on the instance-level SOC test set which contains 600 images. The top of histogram shows the accuracy of AP metric and the bottom of histogram shows the accuracy of AP$^r$. Overall, each module has an indispensable contribution to the entire framework.

TABLE V: Quantitative comparisons with existing methods by AP and AP$^r$ metrics. - indicates unacquirable value. The experiment is evaluated on Dataset1K test set [7] which contains 300 images. For a fair comparison, both our method and S4Net [15] use ResNet-101 as backbone. We keep the rest part of the framework in line.

| Methods | AP | AP$^r$ | AP$^r$ | AP$^r$ | AP$^r$ |
| --- | --- | --- | --- | --- | --- |
| MSRNet [7] | - | 65.32 | - | 52.18 | - |
| S4Net [15] | 52.58 | 86.62 | 78.8 | 63.72 | 41.8 |
| CGCNet | 57.13 | 85.60 | 78.86 | 71.02 | 53.31 |

 aren't available for the related code of [7]. It is inferior to S4Net and the proposed method in performance. ..
Fig. 6: Representative experimental results for each attribute produced by S4Net and the proposed method. Both frameworks are fine-tuned on the Dataset1k training set [7] and tested on the SOC test set [57]. We select a most representative sample in each attribute-based test subset. Each row displays one attribute. We keep the setting of two framework in line.
Fig. 7: Example of failure modes generated by our method. Samples are selected from the Dataset1k test set.

Introduced a new SOD dataset called SOC, which contains both binary mask and instance-level salient ground-truth. Considering that the dataset labels salient instances in clutter, the difficulty of input images is relatively high and so that the experiment results testing on the SOC will be lower than other datasets. In this experiment, we analyze the proposed CGCNet based on image attributes on the test set of SOC dataset. According to the characteristics description in SOC, the instance-level test set is divided into nine attributes: Appearance Change (AC), Big Object (BO), Clutter (CL), Heterogeneous Object (HO), Motion Blur (MB), Occlusion (OC), Out-of-View (OV), Shape Complexity (SC) and Small Object (SO). Each attribute subset has some complexities that need to be overcome. We compared the experimental results of S4Net on the SOC dataset according to several attributes. In order to ensure fairness, both methods are trained on Dataset1K training set, and then directly tested on the SOC test set. Other settings both in two methods are as consistent as possible. The histograms in Fig. 5 show the performance of the CGCNet and S4Net on different attribute test subsets. It can be seen that the proposed method completely outperforms S4Net in AP values, although it is relatively close in terms of AP0.5 value. The advantage may be attributed to the better suppression of complex background by the GFR module. The right histogram demonstrates that the proposed method is more generalized for images with various complex characteristics.

Fig. 6 exhibits typical inferences made by S4Net and our framework according to different attributes. Compared to the Dataset1K, the test set in SOC contains more different kinds of images and the complexity of background is higher. Thanks to the good generation of the CGCNet, our method also shows great performance on the SOC dataset against the S4Net. For example, the sample in first row has the obvious illumination change in salient instance area combining with messy background, the proposed method can easily dig out salient instances from background. The Clutter-based (CL) image has several small salient instances, and the foreground and background regions around instances have similar color. The CGCNet can still accurately locate each instance and handle them calmly. Refer to the last two rows of Fig. 6, instances in images with SC and SO attributes have complex boundaries and are relatively small. Although it is not easy to split the slender legs of the giraffe, the overall result is as expected.

Limitations: Fig. 7 displays some typical failure cases. According to the first row, our method is insensitive to the tenuous local features. Due to the two-stage framework, it is not efficient to suppress the number of proposals in the second row. This strategy can easily lead to more predicted salient instances than the ground-truth. The third row shows that when two salient instances overlap, the detail of boundary is terrible. This is because our weak annotation is composed of bounding boxes and salient regions, which will cause the edge of the salient instance to be the edge of boxes. The bottom two cases demonstrate that our approach fails to predict the salient regions. The problem is very common in saliency detection tasks. In general, it is benefit to use our weak labels based on the proposed CGCNet.

V. Conclusion

In this paper, we present an end-to-end cyclic global context neural network (CGCNet) for salient instance segmentation. Due to lack of dataset for the new challenging task, we used weakly supervised method to train our framework for the first time. Besides, we contributed a weak annotation dataset which contains 4,500 training set and 500 validation set. More importantly, adding with the GFR module and the updating scheme in CGCNet, our framework shows excellent performance for salient instance segmentation, which compares favorably against even fully supervised methods. Due to depend on the region proposal method, the framework sometimes predicts the number of salient instances inaccurately. In the future work, we will further improve the performance of the framework and attempt to exploit one-stage single network for segmenting salient instances.
REFERENCES

[1] J. Lei, B. Wang, Y. Fang, W. Lin, P. L. Callet, N. Ling, and C. Hou, “A universal framework for salient object detection,” IEEE Transactions on Multimedia, vol. 19, no. 9, pp. 1763–1768, 2017.

[2] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, “Learning to detect a salient object,” IEEE Transactions on Pattern analysis and machine intelligence, vol. 33, no. 2, pp. 353–367, 2010.

[3] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, “Perceptnet: Edge guidance network for salient object detection,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8779–8788.

[4] S. Chen, X. Tan, B. Wang, and X. Hu, “Reverse attention for salient object detection,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 234–250.

[5] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, “A generic framework of user attention model and its application in video summarization,” IEEE transactions on multimedia, vol. 7, no. 3, pp. 907–919, 2005.

[6] Y. Gao, M. Shi, J. Tao, and C. Xu, “Detect saliency for fast image retrieval,” IEEE Transactions on Multimedia, vol. 17, no. 3, pp. 359–369, 2015.

[7] G. Li, Y. Xie, L. Lin, and Y. Yu, “Instance-level salient object segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2386–2395.

[8] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, “Deeple3: Deep instance co-segmentation by co-peak search and co-saliency detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8846–8855.

[9] H. Xiao, J. Feng, Y. Wei, M. Zhang, and S. Yan, “Deep salient object detection with dense connections and distraction diagnosis,” IEEE Transactions on Multimedia, vol. 20, no. 12, pp. 3239–3251, 2018.

[10] G. Li and Y. Yu, “Visual saliency based on multiscale deep features,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5455–5463.

[11] J. Dai, Y. Li, K. He, and J. Sun, “R-fcn: Object detection via region-based fully convolutional networks,” in Advances in neural information processing systems, 2016, pp. 379–387.

[12] D. A. Klein and S. Frintrop, “Center-surround divergence of feature statistics for salient object detection,” in 2011 International Conference on Computer Vision. IEEE, 2011, pp. 2214–2219.

[13] C. Xia, F. Qi, and G. Shi, “Bottom-up visual saliency estimation with deep autoencoder-based sparse reconstruction,” IEEE transactions on neural networks and learning systems, vol. 27, no. 6, pp. 1227–1240, 2016.

[14] L. Ye, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, “Learning to detect a salient object,” IEEE Transactions on Pattern analysis and machine intelligence, vol. 33, no. 2, pp. 353–367, 2010.

[15] R. Fan, M.-M. Cheng, Q. Hou, A. Borji, Z. Tu, and P. H. Torr, “Deeply supervised salient object detection with short connections,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3203–3212.

[16] J. Dai, Y. Li, K. He, and J. Sun, “R-fcn: Object detection via region-based fully convolutional networks,” in Advances in neural information processing systems, 2016, pp. 379–387.

[17] Y. Zhu, Y. Zhou, H. Xu, Q. Ye, D. Doermann, and J. Jiao, “Learning instance activation maps for weakly supervised instance segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3116–3125.

[18] R. Fan, M.-M. Cheng, Q. Hou, A. Borji, Z. Tu, and P. H. Torr, “Deeply supervised salient object detection with short connections,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3203–3212.

[19] X. Qin, Z. Zhang, C. Cao, M. Dehghan, and M. Jagersand, “Basnet: Boundary-aware salient object detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7479–7489.

[20] Y. Lee and J. Park, “Centermask: Real-time anchor-free instance segmentation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13906–13915.

[21] R. Fan, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, “Deeply supervised salient object detection with short connections,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1203–1220.

[22] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, “Blendmask: Top-down meets bottom-up instance segmentation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8573–8581.

[23] X. Zhou, Z. Zhou, X. Ruan, Z. Liu, Y. Ye, and J. Xiao, “Learning instance activation maps for weakly supervised instance segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3116–3125.

[24] H. Bilen and A. Vedaldi, “Weakly supervised deep detection networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2846–2854.

[25] S. Huo, Y. Zhou, J. Verbeek, and C. Schmid, “Weakly supervised object localization with multi-fold multiple instance learning,” IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 1, pp. 189–203, 2016.

[26] A. Diba, V. Sharma, A. Pazandesht, H. Pirsiavash, and L. Van Gool, “Weakly supervised cascaded convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 914–922.

[27] P. Tang, X. Wang, A. Wang, Y. Yan, W. Liu, J. Huang, and A. Yuille, “Weakly supervised region proposal network and object detection,” in Proceedings of the European Conference on computer vision (ECCV), 2018, pp. 352–368.

[28] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, “Exploiting saliency for object segmentation from image level labels,” in 2017 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, 2017, pp. 5038–5047.

[29] G. Li, Y. Xie, and L. Lin, “Weakly supervised salient object detection using image labels,” in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[30] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, “Learning to detect salient objects with image-level supervision,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 136–145.

[31] Y. Zeng, Y. Zhuge, H. Liu, L. Zhang, M. Qian, and Y. Yu, “Multi-source weak supervision for saliency detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6074–6083.

[32] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[45] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.

[46] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in Advances in neural information processing systems, 2015, pp. 91–99.

[47] A. Neubeck and L. Van Gool, “Efficient non-maximum suppression,” in 18th International Conference on Pattern Recognition (ICPR’06), vol. 3. IEEE, 2006, pp. 850–855.

[48] J. Zhang, S. Ma, M. Samek, S. Sclaroff, M. Betke, Z. Lin, X. Shen, B. Price, and R. Mech, “Salient object subitizing,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4045–4054.

[49] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807–814.

[50] J. Shotton, J. Winn, C. Rother, and A. Criminisi, “Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context,” International Journal of Computer Vision, vol. 81, no. 1, pp. 2–23, 2009.

[51] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, “Predicting human eye fixations via an Istm-based saliency attentive model,” IEEE Transactions on Image Processing, vol. 27, no. 10, pp. 5142–5154, 2018.

[52] R. Girshick, “Fast r-cnn,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[54] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, “Simultaneous detection and segmentation,” in European Conference on Computer Vision. Springer, 2014, pp. 297–312.

[55] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[56] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.

[57] 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 European Conference on Computer Vision (ECCV). Springer, 2018.