Salient Instance Segmentation with Region and Box-level Annotations

Jialun Pei, He Tang, Tianyang Cheng and Chuanbo Chen

Abstract—**Salient instance segmentation is a new challenging task that received widespread attention in the saliency detection area. The new generation of saliency detection provides a strong theoretical and technical basis for video surveillance. Due to the limited scale of the existing dataset and the high mask annotations cost, plenty of supervision source is urgently needed to train a well-performing salient instance model. In this paper, we aim to train a novel salient instance segmentation framework by an inexact supervision without resorting to laborious labeling. To this end, we present a cyclic global context salient instance segmentation network (CGCNet), which is supervised by the combination of salient regions and bounding boxes from the ready-made salient object detection datasets. To locate salient instance more accurately, a global feature refining layer is proposed that dilates the features of the region of interest (ROI) to the global context in a scene. Meanwhile, a labeling updating scheme is embedded in the proposed framework to update the coarse-grained labels for next iteration. Experiment results demonstrate that the proposed end-to-end framework trained by inexact 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 58.3% in the test set of Dataset1K that outperforms the mainstream state-of-the-art methods.**

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

I. INTRODUCTION

S ALIENT object detection (SOD) is known as a classic research field for highlighting the most sensitive and informative regions in a scene [1]–[3]. Originating from cognitive and psychology research communities, salient object detection is applied to various areas, such as video surveillance [4], video summarization [5] and content-aware image editing [6]. With the rapid development of current image acquisition equipment and 5G communication technology, the traditional binary mask of salient object detection is inadequate to meet the needs of high-resolution image segmentation. Albeit salient object detection task provides the salient region labels compared to the background, they do not explore instance-level cue for 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 concentrate on the new challenging task salient instance segmentation (SIS) for improving the intelligence level of monitoring systems.

Visual saliency 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 the computer vision area [12], [13], especially focusing on estimating the bounding boxes of salient instances [14]. Different from salient object detection, salient instance segmentation fosters a more detailed information by labeling each instance with a precise pixel-wise mask and promotes the saliency maps from region-level to instance-level for more detailed analysis. In contrast to instance segmentation, salient instance segmentation only predicts salient instances based on the salient regions. Moreover, segmenting salient instances is class-agnostic compared to the class-specific instance segmentation task.

However, the saliency models of CNNs are usually required to the pixel-level fully-supervised train data [15], [16]. Up to now, the existing SIS dataset is seriously inadequate and the amount of pixel-wise ground-truths is insufficient in a single dataset. The quality and quantity of pixel-level annotations is the bottleneck because the labeling task is strenuous and time-consuming. To alleviate the effectiveness of lacking fully-supervised data, weakly supervised learning is viewed as the alternative training method attracting more attention. This strategy not only avoids user-intensive labeling, but also encourages the models to receive enough training samples.

Inspired by this consideration, in this paper, we aim to integrate the bounding boxes and binary salient regions for training the SIS frameworks. The bounding box annotation contains location information for each salient instance. Meanwhile, salient regions provide salient region information which is a ready-made source generated from the existing SOD datasets. Both box-level and region-level annotations are inexact for salient instance segmentation [17]. As shown in Fig. 1, the bounding boxes determine the location and number of salient instances which have labeled in the DUT-OMRON dataset [18]. We use the bounding box and salient region to assign salient regions to each bounding box of salient instance. It is essential to combine these two supervision sources because the bounding box annotation lacks the pixel-level labels and salient region cannot distinguish different salient instances in the coarse-grained labels. To ensure one instance corresponds to one bounding box and hold the consistency of salient instances and regions, we also exploit some priors to prevent the different object regions trapped into the same box. In this case, the network can utilize more training samples with the lowest labeling cost. We will elaborate the generation steps of the coarse-grained annotations in Section III.C.

* He Tang is the corresponding author.

Jialun Pei is with the School of Computer Science and Technology, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan, 430074, China (e-mail: peijl@hust.edu.cn)

He Tang, Tianyang Cheng and Chuanbo Chen are with the School of Software Engineering, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan, 430074, China (e-mail: hetang@hust.edu.cn; patrickcty@hust.edu.cn; chuanboc@163.com)

| (a) Input image | (b) Box and region-level annotation | (c) Weak annotations | (d) Our salient instance result |

Fig. 1: The coarse-grained annotation is generated to achieve salient instance results by the proposed framework. (b) shows the combination of bounding box and salient region annotations. (c) exhibits the coarse-grained labels for inexact supervised learning. The final result predicted by CGCNet is showed in (d).

For segmenting salient instances, we design a cyclic global context SIS network (CGCNet) supervised by the above coarse-grained labels. Fig. 2 shows the overview of our CGCNet. The proposed model is an end-to-end two-stage SIS framework, which first detects salient proposals and then predict the pixel-level salient instance masks. When extracting features for salient mask prediction, the performance of convolutional layer depends heavily on global context. Considering obtaining stronger feature representation, we extend the scope of feature extraction from the local proposal to the global features. Inspired by enter-surround contrast derived from saliency detection mechanism [19]–[21], a global feature refining module (GFR) is designed to make full use of background features and suppress disturbance from other salient instance features [22]. Different from the ROIAlign layer that limits the receptive field in Mask R-CNN [23], the proposed GFR module is sensitive to global contrast in order to capture more detailed edge information. Moreover, the CGCNet is designed to iteratively update the coarse-grained annotations by using the forward prediction masks combining with a conditional random field (CRF) [24]. It is beneficial to refine the coarse-grained annotations sequentially. The input training samples and the corresponding results are shown in Fig. 1. We evaluate the results on the test set of Dataset1K [7] and show that our method compares favourably against even some fully supervised methods.

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

- We propose a novel inexact supervision salient instance segmentation framework called cyclic global context network (CGCNet), which is supervised by the combination of the region-level bounding boxes and salient regions.
- We design a global feature refining (GFR) layer that extends the receptive field of each instance to the global context and suppress the features of other salient instances simultaneously.
- We embed an update scheme in CGCNet that can optimize the coarse-grained labels continuously to improve the accuracy.

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 comparions 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 performance of SOD models boost significantly. Fortunately, rich pixel-level salient datasets can be poured into various CNN models to detect salient regions [25], [26]. Li *et al.* [27] proposed a multi-scale deep contrast network to overcome the limitations of overlap and redundancy. Hou *et al.* [28] designed short connections to the skip-layer structures based on the VGGNet for better supervision. Qin *et al.* [29] produced 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 viewed as the region-level binary mask which may not accomplish instance-level salient object segmentation.

B. Salient Instance Segmentation

Proceed from SOD, salient instance segmentation propels the problem into an instance-level phase. Unlike instance segmentation [30]–[32], salient instance is category-independent and it is concentrate on salient regions. Therefore, the frameworks and datasets of instance segmentation are incompatible with segmenting salient instances. 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 to raise saliency detection from the region level to the instance level. Subsequently, Li *et al.* [7]

Fig. 2: An overview of the proposed framework. The detail of the GFR module is shown in the upper right corner. The coarse-grained annotations updating criteria is illustrated in Section III.D. At the training time, the salient instance result return to update the coarse-grained annotation in next iteration.

formally proposed the instance-level salient object detection task. They drove the prediction results from proposals to pixel-level, and produced the first SIS dataset containing 1,000 samples. Pei *et al.* [33] proposed a multi-task model to predict salient regions and subitizing, and then applied a spectral clustering algorithm to segment salient instances. Recently, Fan *et al.* [34] proposed an end-to-end single-shot salient instance segmentation framework to segment salient instances. The proposed ROIMasking layer allows more detailed information to be detected accurately, and meanwhile remains the context information around the regions of interest. As a new challenging task, however, the lacking of fully-supervised label is the main problem to limit the performance of deep learning models. To avoid making the high cost of pixel-level annotations, we take advantage of the inexact supervision to train our model.

C. Weakly Supervised Learning

Most neural networks require full 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-labelling, weakly supervised learning has attracted a great deal of attention in recent years [36]–[38]. Many weakly supervised principles have been introduced in machine vision area, including object detection, instance segmentation and saliency detection [39], [40]. Weakly supervised learning reveals that the 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 un-supervised methods as pixel-level annotation to detect salient objects. Zheng *et al.* [42] take advantage of salient subitizing as the weak supervision to generate the initial saliency maps,

and then propose a saliency updating module (SUM) to refine the saliency maps iteratively. 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 build an inexact label which embraces the existing binary salient regions and bounding boxes for better training the SIS network.

III. THE CGCNET ARCHITECTURE

A. Motivation

The motivation of the proposed method is handled with seg-menting class-agnostic salient instances under lacking fully-supervised annotations. We tend to utilize sufficient training samples with the lowest labeling cost. Therefore, in this paper, the coarse-grained label is proposed that is composed of bounding boxes and binary salient regions. On one hand, the salient proposals provide positional information of salient instance. On the other hand, binary salient regions can provide approximate salient area information for salient instances. Additionally, they can be easily achieved from existing SOD datasets. For training by the coarse-grained labels, we design a cyclic global context neural network (CGCNet) to predict salient instances and update the coarse-grained labels recur-rently.

B. Overall Framework

As shown in Fig. 2, the framework of our proposed CGCNet consists of three main components. Firstly, The RPN head is viewed as a salient proposal detector that detects the bounding boxes of salient instance to capture the location and number of salient instances. Then, the GFR module provides the global feature representation to predict salient masks.

Moreover, the resulting salient instances update the coarse-grained ground-truth added with the fully connected CRF operation for the next iteration.

We combine pre-trained ResNet-101 [44] with FPN [45] as the backbone. According to the order of downsampling in ResNet-101, we extract the 4-th stage feature map followed by a 1×1 convolutional layer with the lateral connections in multi-level FPN prediction [23]. Followed by FPN, we utilize five levels of feature maps to detect different sizes of objects on different levels to maximize the gains in accuracy. The feature maps produced by the backbone are extracted from the entire input image. Both salient proposal detector and salient instance segmentation branch are 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 the category-independent characteristic, each ROI feature is assigned to two classes, denoted as $B_c(c \in \{0,1\})$. The two classifications correspond to the background and the salient object in foreground. 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 embedded to suppress the negative proposals that 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 phase, 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 features processed by the GFR module are injected into a pixel-to-pixel fully convolutional block. The Fully convolutional fashion preserves the spatial consistency of each pixel involved in corresponding salient instances. Moreover, taking the resulting salient instances predicted by the SIS branch, the updating scheme is produced to update the coarse-grained ground-truth recurrently in training phase. In the following subsection, we will describe the SIS branch and the GFR module in detail.

C. Inexact Supervision Sources

We implement the coarse-grained annotations to handle the problem of lacking sufficient labels for the SIS task. Considering the characteristics of salient instances, it is essential to embrace both the salient region and the number of salient instances. Inspired by salient object detection and instance segmentation tasks, the coarse-grained labels are composed of salient regions and the bounding boxes of salient instances.

To train the proposed CGCNet model, we utilize the largest number of SOD dataset called DUT-OMRON [18], which contains about 5,000 salient object labels and the bounding boxes. We select 4,500 images from the training set of the DUT-OMRON SOD dataset. Despite combining salient regions and bounding boxes, the coarse-grained labels still have some general issues. First, salient regions from different bounding boxes have shared patches. Second, 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 bounding box corresponding to each salient instance is denoted as $W_i(i = 1, 2, \ldots, n)$. In addition, we set the patches discarded by priors in each window as φ_i. The final coarse-grained label I is defined by:

$$I = \sum_{i=1}^{n} [S(x,y) \cap W_i - \varphi_i(\hat{x}, \hat{y})], \quad i = 1, 2, ..., n, \quad (1)$$

where (x, y) presents salient region pixels in the image S and $(\hat{x}, \hat{y})$ denotes the set of pixels excluded by our priors in each window. n is the number of salient instances in an image. 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. By achieving the ROI features 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 segmentation. Inspired by center-surround contrast based on the SOD task, we seek to extend the ROI feature to the global feature map. Resorting to increasing receptive field and ensuring the resolution of instances, we utilize global features extracted from the backbone instead of the ROI feature. Meanwhile, each feature map produced from the GFR module only contains the feature of current salient instance proposal and background while suppressing the features of other salient instance proposals.

The GFR module. The goal of the proposed global feature refining module (GFR) is to obtain global context information and limit the disturbance of other instance features.For the ROIAlign module, it only pay attention to the ROI feature and resize the original resolution of ROI [23]. In S4Net [34], the ROIMasking extend the receptive field and use of the information around the ROI contrasting ROI features. Differ from the ROIMasking, 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 transfers the coordinates of all proposals from different scales of features to the aspect ratio of original feature map. Tasking $F^{(H \times W \times C)}$ as the input feature map, we assume that the number of proposals is n. To explain the module more facilitatively, the number of proposals is set to 3. Let $R_i^{(H \times W \times C)}(i = 1, 2, \ldots, n)$ as the feature map includes i-th features of proposals. To maintain the consistency of resolution between F and R_i, global average pooling is used to fill in the background area. The output of GFR module $G_i(i = 1, 2, \ldots, n)$ is defined by:

$$G_i = F - \sum_{i=1}^{n} R_i + R_i, \quad i = 1, 2, ..., n \quad (2)$$

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].

Each feature map G_i contains the corresponding feature of proposals and the feature of background. To constrain other features of proposals, the operation of our GFR module first digs out all regions of salient proposals in input feature map and then sticks the corresponding ROI feature on F according to the coordinates of proposal. This operation also avoids missing the shared pixels from different proposals and reserves the occlusion parts.

Fig. 3 visualizes the process of GFR module and compares with other similar modules. We also introduce the local feature refining module (LFR). Compared to the GFR module, the LFR module extends the receptive field based on the ROI feature while limiting other salient proposal features rather than covering global features. Assume that the size of salient proposal is (H_r, W_r), the size of extended bounding box is set to $(H_r + h, W_r + w)$, where h and w is $H_r/5$ and $W_r/5$, respectively. The other setting of LFR module is same as GFR module. Additionally, the corresponding process in Mask R-CNN [23] is exhibited in the top branch in Fig. 3. The experiment results demonstrate that the GFR module outperforms the other two modules for SIS task, which is discussed in detail in Section IV.C.

After adding with our GFR module, each target instance not only contains the features inside the proposal but also takes advantage of the global context information to highlight the instance region. The mask head is efficient to 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 [48]. All the convolutional layers have a kernel size 3×3 and stride 1.

Coarse-grained Annotations Updating Scheme. Considering the initial training samples are coarse-grained annotations, we produce an updating scheme to optimize coarse-grained annotations continuously. The fundamental flaw of the coarse-grained labels is that boundary information of each instance is not detailed enough, and different instances in one

image have overlap and occlusion. If only training on the original samples, the predicted salient instances would contain some small redundant patches that 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:

$$E(M) = -\sum_i log P(m_i) + \sum_{i,j} \varphi_p(m_i, m_j), \quad (3)$$

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:

$$\varphi_p(m_i, m_j) = \omega_1 \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right) + \\ w_2 \exp\left(-\frac{|p_i - p_j|}{2\theta_\gamma^2}\right)^2, \quad (4)$$

where the first kernel depends on pixel positions p and pixel intensities I. The kernel encourages nearby pixels with similar features to take consistent salient instance labels [27]. The second kernel quantifies the smoothness kernel which only depends on pixel positions for removing small isolated regions [49]. ω_1 and ω_2 indicate the weighted values to balance the two parts. The hyper parameters θ_α, θ_β and θ_γ 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 ω_1, ω_2, θ_α, θ_β and θ_γ 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 ω_2 and θ_γ are set to 3 and 1, and the initial search range of the parameters are $\omega_1 \in [1:1:10,$

$\theta_\alpha \in [50:5:100]$ and $\theta_\beta \in [5:1:15]$. These parameters are fixed through 10 iterations of the average field to achieve the best value. In our experiments, the values of ω_1, ω_2, θ_α, θ_β and θ_γ are set to 4, 3, 70, 13, 1, respectively.

We denote the salient instance map as R and the map processed by CRF as R_f. The coarse-grained annotation is labeled as C. According to Algorithm 1, we propose a strategy based on the KL-Divergence [50] to update the C for the next iteration. KL-Divergence is defined as a 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 CGCNet is noisy, the updating prediction map should have more dissimilar patches with coarse-grained annotation as well as the larger value of KL-Divergence between them. Our strategy compares the prediction map R and R_f to the coarse-grained annotation C, which is designed as:

$$K_1(R, C) = \frac{1}{H \times W} \sum_{i=1}^{H \times W} C_i log(\frac{C_i}{R_i + \sigma} + \sigma) \quad (5)$$

$$K_2(R_f, C) = \frac{1}{H \times W} \sum_{i=1}^{H \times W} C_i log(\frac{C_i}{R_{f_i} + \sigma} + \sigma), \quad (6)$$

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 σ is a regularization constant. In Algorithm 1, C_n represents the ground-truth to be used for the next iteration. It is observed that we set φ as the threshold to determine whether to update with the existing coarse-grained annotation C. The value of φ is set to 0.05. The strategy can eliminate redundant replacements and alleviate the impact of excessive erosion of the CRF on the prediction map. Using the updating scheme to the inexact supervised learning, the network achieved more accurate results at the training phase.

Algorithm 1 Coarse-grained annotations updating

Input: Coarse-grained annotation C, salient instance map R and salient instance map with CRF R_f.

Ensure: The updated coarse-grained 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**

Loss Function. The proposed CGCNet need to trained salient proposal branch and SIS branch simultaneously. Therefore, we use ground-truth proposals to supervise the RPN head and the pixel-level coarse-grained labels to train SIS branch. The loss function of CGCNet is defined as a two-stage fashion:

$$L = L_{bb} + L_{seg} + L_{upd} \quad (7)$$

Where 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 [51]. The SIS branch loss L_{seg} is defined by the cross-entropy loss, which is followed by:

$$L_{seg} = -\frac{1}{N} \sum_{i=1}^{N} (g_i log p_i + (1 - g_i) log(1 - p_i)) \quad (8)$$

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:

$$L_{upd} = K_2(R_f, C) - K_1(R, C) \quad (9)$$

In the training phase, the weights of the backbone are frozen. The entire procedure is repeated iteratively for training.

IV. EXPERIMENTAL RESULTS

In this section, we elaborate on the results of the proposed CGCNet framework for the SIS task in detail. We perform ablation experiments on various components of our approach. Besides, we use different metrics to compare with the experimental results of other state-of-the-art methods. Since the proposed method accomplishes the SIS task by inexact supervised learning, we will maintain maximum fairness in comparison.

A. Implementation Details

As described in the section above, the end-to-end CGCNet is trained by our inexact labels which select 4,500 images from DUT-OMRON dataset [18] without ambiguous samples. During training, the salient bounding box ground-truths are used to supervise the salient proposal detector while combining with SOD annotations to train the mask branch. Meanwhile, we utilize 500 images as the same as training data for validation. For training salient proposals, the bounding boxes are considered as a positive sample if the IOU is more than 0.7 or a negative sample below 0.3. In addition, the NMS threshold used on the proposal detector is set to 0.7. At inference time, we only use 300 images from the testing set in the dataset proposed in [7] due to shortage of datasets. We input the number of top 80 scoring proposals from the proposal prediction branch after applying NMS to the GFR module. Additionally, the SIS branch directly outputs 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. To speed up training convergence, we initialize the CGCNet with a pre-trained model over the ImageNet dataset [52] from Mask R-CNN [23]. 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 brand new task, salient instance segmentation has few evaluation metrics to measure its performance quantitatively. Different from SOD and instance segmentation, The SIS task distinguishes pixel-level instances based on salient regions without classification. Therefore, 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 [56]. The

TABLE I: Comparison of different backbones used in the CGCNet on DUT-ORMON validation set. In this experiment, we keep the rest part of the framework in line.

Backbone	AP	$AP^r0.5$	$AP^r0.7$
VGG16 [53]	50.79	79.28	60.38
ResNet-50 [41]	57.13	85.6	71.02
ResNet-101 [54]	57.69	86.04	71.72
ResNeXt-101 [55]	**58.28**	**86.91**	**72.69**

TABLE II: Ablation study for different modules in SIS branch. The experiment is evaluated on DUT-ORMON validation set.

Modules	LFR module	GFR module	ROIAlign [23]	ROIMasking [34]
$AP^r0.5$	85.45	**86.04**	85.25	85.73
$AP^r0.7$	70.2	**71.72**	70.28	70.46

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^r metric is defined by the summation of precision value divided by the number of all images in testing set, which is formulated as:

$$AP^r\alpha = \frac{1}{N}\sum_j \frac{1}{n}\sum_i precision, \quad IoU(i) \geq \alpha \quad (10)$$

$$precision = \begin{cases} 1, & if\ IoU(i) \geq \alpha \\ 0, & if\ IoU(i) < \alpha \end{cases}, \quad (11)$$

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

$$AP = \frac{1}{10}\sum_\alpha AP^r|\alpha, \quad \alpha = 0.5, 0.55, ..., 0.95 \quad (12)$$

Compared with the AP^r metric, the AP value is adopted to measure the overall performance of SIS methods. In this section, the experimental results are evaluated mainly based on the above-mentioned two metrics.

C. Ablation Studies

We analyze the effectiveness of the proposed CGCNet on DUT-OMRON validation set [18]. The ablation studies contain four parts: performance of four different backbones, performance of GFR module versus three related structures, hyper-parameter of the updating scheme and contributions of each component of our framework.

Backbone: To ensure fairness and the effects of the different backbones on the experimental results, we verify various backbones working on CGCNet which stay in the 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^r metric [55]. The widely used ResNet-101 has also achieved good results slightly behind ResNeXt-101.

TABLE III: The threshold φ of updating scheme performance of CGCNet. The highest scores in each row are labeled in bold.

φ	0.01	0.05	0.1	0.15	0.2
$AP^r0.5$	85.89	**86.04**	84.85	84.66	84.13
$AP^r0.7$	71.34	**71.72**	71.16	70.68	70.19

Due to insufficient depth of the network, VGGNet obtained relatively low accuracy, but is slightly faster than ResNet [53].

The GFR Module: The proposed GFR module is viewed as the core layer in SIS branch to refine features. In this section, we try to evaluate 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 ROIAlign in Mask R-CNN [23] and ROIMasking in S4Net [34]. As shown in Table. II, the experimental results based on GFR module outperforms other modules. ROIAlign only concentrates on the ROI features. Albeit the LFR module extended the scale of features around ROI, it still slightly behind the ROIMasking by reason of its ternary masking. It indicates that treatment of refining features play an important role in segmenting salient instances. Finally, we adopt the GFR module embedded in our framework.

Hyper-parameter in updating scheme: The threshold φ of updating scheme is essential for the quality of inexact 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 [24], we empirically provide several default values for determining its influence in this experiment, which is shown in Table. III. The performance of different values of φ is relatively average. The best result is obtained when the value of φ was set to 0.05, it can balance the optimal quantity and quality of replacement.

The component in CGCNet: We conducted extensive experiments to discover contributions of each innovative module under the same settings. These parts of CGCNet include the prior criteria (Standardized coarse-grained 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 that improved the AP metric about 2 percent compared to without it. It can be attributed to the insertion of CRF and the revision of the coarse-grained annotations at the training time. With the help of the prior criteria, the performance significantly improved in terms of $AP^r0.5$ and $AP^r0.7$ metrics. Overall, each module has an indispensable contribution to the entire framework.

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

There are three existing methods related to the salient instance segmentation task: MSRNet [7], S4Net [34] and SCNet [33]. In contrast to these previous works, we are the first to make use of inexact supervised learning for the new challenging task. All methods are evaluated on the test set of Dataset1K [7] and SOC dataset [57], respectively. For fair comparison, we compare the existing salient instance

Input	Saliency Map	Ground-truth	S4Net	CGCNet

Fig. 4: Qualitative analysis of experimental results by the proposed method and S4Net [34].

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. Considering that all existing methods are fully supervised and our method is supervised by inexact labels, we train all methods on the training set of Dataset1K and our coarse-grained annotations of DUT-OMRON dataset, respectively.

Then, we evaluate these models on the test set of Dataset1K [7]. Since our inexact labels are not applicable to MSRNet and SCNet, we only compare with S4Net by using inexact labels for training. The proposed CGCNet use ResNet-50 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, $AP^r0.5$ and $AP^r0.7$ metric achieved by different training set. Due to the related code of [7] is not available, we cannot obtain its whole results.

Fig. 5: The attributes-based performance of the CGCNet on the instance-level SOC test set. The left of histogram shows the accuracy of AP metric. The histograms in the middle and right show the accuracy of $AP^r0.5$ and $AP^r0.7$ metric under nine attributes.

TABLE IV: Ablation analysis of effects 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-ORMON validation set.

Models	AP	$AP^r0.5$	$AP^r0.7$
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	**57.69**	**86.04**	**71.72**

TABLE V: Quantitative comparisons with existing methods on the training set of our inexact labels and Dataset1K [7], respectively. The results are evaluated on the test set of Dataset1K [7]. For a fair comparison, both our method and S4Net [34] use ResNet-50 as backbone. We keep the rest part of the framework in line. '-' indicates unacquirable value.

Method	Training Set	AP	$AP^r0.5$	$AP^r0.7$
S4Net [34]	DUT-ORMON	50.9	84.9	60.8
CGCNet (Ours)	(Inexact labels)	**58.3**	**88.4**	**71.0**
MSRNet [7]		-	65.3	52.3
SCNet [33]	Dataset1K [7]	56.8	84.6	67.4
S4Net [34]		52.3	**86.7**	63.6
CGCNet (Ours)		**57.1**	85.8	**69.0**

In the case of training on the inexact labels, our method achieves the best result compared to all other methods. As an inexact supervised method, the CGCNet improves the value of AP metric to the highest 58.3%. Additionally, we also exhibit the results of our framework and other fully-supervised methods on the training set of Dataset1K [7]. As shown in the bottom of Table. V, the value of AP achieved by our CGCNet also outperforms SCNet and S4Net. While the value of $AP^r0.5$ metric is slightly lower than S4Net, our framework has demonstrated its robustness whether trained on inexact labels or not.

We also qualitatively analyzed the experimental results produced by CGCNet and S4Net. Fig. 4 displays some results from the testing set in Dataset1K [7]. It shows 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. Our 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 also distinguish obstructed instances satisfactorily. In comparison, the S4Net determine the number of salient instances inaccurately in some cases. The antepenult sample demonstrated that the S4Net is insensitive to smaller salient instances. 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 performance of S4Net. By and large, the proposed framework has high accuracy and robustness for salient instance segmentation.

Evaluation on the SOC: Recently, Fan *et al.* [57] introduced a Salient Object in Clutter 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. Therefore, the experiment results will be lower than other datasets. In this experiment, we analyze the proposed CGCNet in terms of image attributes on the test set of SOC dataset. 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) [57]. We compare the experimental results of S4Net according to the attributes. For fair comparison, both methods are trained on Dataset1K training set [7], and then directly tested on the SOC test set. The histograms in Fig. 5 show the performance of the CGCNet and S4Net on different attribute test subsets. Although these two methods achieve approximate scores in terms of $AP^r0.5$ metric, CGCNet performs significantly in AP values. It can 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 different attributes. Moreover, our framework excels at dealing with the image containing heterogeneous object (HO) compared to other attributes. Thanks to the global features of the GFR module, CGCNet process the image with AC attribute effectively. The AP value of OC attribute is lowest because the occluded part of object is difficult to detect. Overall, our method is robust for processing images with different attributes.

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.

Input	Ground-truth	CGCNet

Fig. 7: Example of failure modes generated by our method. Samples are selected from the Dataset1k test set [7]

Fig. 6 exhibits some typical results generated 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. Our method also shows great performance on the SOC dataset against the S4Net. For example, the sample in the 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 proposed CGCNet can still accurately locate each instance and segment them out. Refer to the last two rows of Fig. 6, salient 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 satisfying.

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 inefficient to suppress the number of proposals in the second row. This strategy tends to result in a greater number of predicted salient instances than the ground-truth. The third row shows that the detail of the boundary is terrible when two salient instances overlap. It is due to the inexact annotations consisting of bounding boxes and salient regions that cause the edge of the salient instance to become 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. Generally, it is beneficial to use coarse-grained labels based on the proposed CGCNet.

V. CONCLUSION

In this paper, we propose 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 inexact supervised learning to train our framework. 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 some fully supervised methods. Due to dependence on the post processing of NMS, the framework sometimes predicts the number of salient instances inaccurately. In the future work, we will attempt to exploit one-stage single network and further improve the effectiveness of the framework for applying to video surveillance.

REFERENCES

[1] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 8, pp. 1309–1321, 2015.

[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, "Egnet: Edge guidance network for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8779–8788.

[4] Z. Shao, L. Wang, Z. Wang, W. Du, and W. Wu, "Saliency-aware convolution neural network for ship detection in surveillance video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 3, pp. 781–794, 2020.

[5] M. Paul and M. Musfequs Salehin, "Spatial and motion saliency prediction method using eye tracker data for video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 6, pp. 1856–1867, 2019.

[6] Y. Gao, M. Shi, D. Tao, and C. Xu, "Database 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, "Deepco3: 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] Z. Tu, Y. Ma, C. Li, J. Tang, and B. Luo, "Edge-guided non-local fully convolutional network for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 582–593, 2021.

[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] F. Guo, W. Wang, Z. Shen, J. Shen, L. Shao, and D. Tao, "Motion-aware rapid video saliency detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4887–4898.

[12] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9627–9636.

[13] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *European Conference on Computer Vision*. Springer, 2016, pp. 75–91.

[14] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Unconstrained salient object detection via proposal subset optimization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5733–5742.

[15] L. Zhu, J. Chen, X. Hu, C.-W. Fu, X. Xu, J. Qin, and P.-A. Heng, "Aggregating attentional dilated features for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3358–3371, 2020.

[16] L. Wang, R. Chen, L. Zhu, H. Xie, and X. Li, "Deep sub-region network for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 728–741, 2021.

[17] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018.

[18] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3166–3173.

[19] 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.

[20] 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.

[21] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 733–740.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[24] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.

[25] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1623–1632.

[26] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Sod-mtgan: Small object detection via multi-task generative adversarial network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 206–221.

[27] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 478–487.

[28] Q. Hou, 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. 3203–3212.

[29] X. Qin, Z. Zhang, C. Huang, C. Gao, 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.

[30] 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. 13 906–13 915.

[31] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo, "Polarmask: Single shot instance segmentation with polar representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 193–12 202.

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

[33] J. Pei, H. Tang, C. Liu, and C. Chen, "Salient instance segmentation via subitizing and clustering," *Neurocomputing*, vol. 402, pp. 423–436, 2020.

[34] R. Fan, M.-M. Cheng, Q. Hou, T.-J. Mu, J. Wang, and S.-M. Hu, "S4net: Single stage salient-instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6103–6112.

[35] 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.

[36] 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.

[37] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE transac-tions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 189–203, 2016.

[38] A. Diba, V. Sharma, A. Pazandeh, 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.

[39] 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.

[40] 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.

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

[42] X. Zheng, X. Tan, J. Zhou, L. Ma, and R. W. Lau, "Weakly-supervised saliency detection via salient object subitizing," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[43] Y. Zeng, Y. Zhuge, H. Lu, 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.

[44] 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] 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.

[49] 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.

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

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

[52] 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.

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

[54] 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.

[55] 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.

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

[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.