Semi-Supervised Video Salient Object Detection Using Pseudo-Labels

Pengxiang Yan[1]   Guanbin Li[1*]   Yuan Xie[1,2]
Zhen Li[3]   Chuan Wang[4]   Tianshui Chen[1,2]   Liang Lin[1,2]
[1]Sun Yat-sen University   [2]DarkMatter AI Research   [4]Megvii Technology
[3]Shenzhen Research Institute of Big Data, the Chinese University of Hong Kong (Shenzhen)

yanpx@mail2.sysu.edu.cn, liguanbin@mail.sysu.edu.cn, xiey39@mail2.sysu.edu.cn,

lizhen36@connect.hku.hk, wangchuan@megvii.com, tianshuichen@gmail.com, linliang@ieee.org

Abstract

Deep learning-based video salient object detection has recently achieved great success with its performance significantly outperforming any other unsupervised methods. However, existing data-driven approaches heavily rely on a large quantity of pixel-wise annotated video frames to deliver such promising results. In this paper, we address the semi-supervised video salient object detection task using pseudo-labels. Specifically, we present an effective video saliency detector that consists of a spatial refinement network and a spatiotemporal module. Based on the same refinement network and motion information in terms of optical flow, we further propose a novel method for generating pixel-level pseudo-labels from sparsely annotated frames. By utilizing the generated pseudo-labels together with a part of manual annotations, our video saliency detector learns spatial and temporal cues for both contrast inference and coherence enhancement, thus producing accurate saliency maps. Experimental results demonstrate that our proposed semi-supervised method even greatly outperforms all the state-of-the-art fully supervised methods across three public benchmarks of VOS, DAVIS, and FBMS.

1. Introduction

Salient object detection aims at identifying the most visually distinctive objects in an image or video that attract human attention. In contrast to the other type of saliency detection, i.e., eye fixation prediction [20, 41] which is designed to locate the focus of human attention, salient object detection focuses on segmenting the most salient objects with precise contours. This topic has drawn widespread interest as it can be applied to a wide range of vision applications, such as object segmentation [46], visual tracking [47], video compression [14], and video summarization [32].
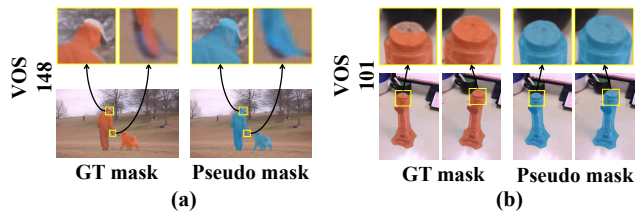
---

*Corresponding author is Guanbin Li.

Figure 1. Example ground truth masks (orange mask) vs. our generated pseudo-labels (blue mask) from the VOS [27] dataset.

Recently, video salient object detection has achieved significant progress [24, 36, 44] due to the development of deep convolutional neural networks (CNNs). However, the performance of these deep learning-based methods comes at the cost of a large quantity of densely annotated frames. It is arduous and time consuming to manually annotate a large number of pixel-level video frames since even an experienced annotator needs several minutes to label a single frame. Moreover, a video clip usually contains hundreds of video frames with similar content. To reduce the impact of label noise on model training, the annotators need to spend considerable time checking the consistency of the label before and after. Considering that visual saliency is subjective, the annotation work becomes even more difficult, and the quality of the labeling is hard to guarantee.

Although there are many unsupervised video salient object detection methods [42, 43, 27] that are free of numerous training samples, these methods suffer from low prediction accuracy and efficiency. Since most of these methods exploit hand-crafted low-level features, e.g., color, gradient or contrast, they work well in some considered cases while failing in other challenging cases. Recent research by Li *et al.* [22] noticed the weakness of unsupervised methods and the lack of annotations for deep learning-based methods. They attempted to use the combination of coarse activation maps and saliency maps, which were generated by learning-based classification networks and unsupervised methods respectively, as pixel-wise training annotations for image salient object detection. However, this method is not suit-

able for the video-based salient object detection task, where object motion and changes in appearance contrast are more attractive to human attention [15] than object categories. Moreover, it is also challenging to train deep learning-based video salient object detection models for temporally consistent saliency map generation, due to the lack of temporal cues in sparsely annotated frames.

By carefully observing the training samples of existing video salient object detection benchmarks [27, 35, 3], we found that the adjacent frames in a video share small differences due to the high video sampling rate (e.g., 24 fps in the DAVIS [35] dataset). Thus, we conjecture that it is not necessary to densely annotate all the frames since some of the annotations can be estimated by exploiting motion information. Moreover, recent work has shown that a well-trained CNN can also correct some manual annotation errors that exist in the training samples [22].

Inspired by these observations, in this paper, we address the semi-supervised video salient object detection task using unannotated frames with pseudo-labels as well as a few sparsely annotated frames. We develop a framework that exploits pixel-wise pseudo-labels generated from a few ground truth labels to train a video-based convolutional network for saliency maps with spatiotemporal coherence. Specifically, we first propose a refinement network with residual connections (RCRNet) to extract spatial saliency information and generate saliency maps with high-resolution through a series of upsampling based refine operations. Then, the RCRNet equipped with a non-locally enhanced recurrent (NER) module is proposed to enhance the spatiotemporal coherence of the resulting saliency maps. For the pseudo-label generation, we adopt a pre-trained FlowNet 2.0 [13] for motion estimation between labeled and unlabeled frames and propagate adjacent labels to unlabeled frames. Meanwhile, another RCRNet is modified to accept multiple channels as input, including RGB channels, propagated adjacent ground truth annotations, and motion estimations, to generate consecutive pixel-wise pseudo-labels, which make up for the temporal information deficiency that exists in sparse annotations. As shown in Fig. 1, our model can produce reasonable and consistent pseudo-labels, which can even improve the boundary details (Example a) and overcome the labeling ambiguity between frames (Example b). Learning under the supervision of generated pseudo-labels together with a few ground truth labels, our proposed RCRNet with NER module (RCRNet+NER) can generate more accurate saliency maps which even outperforms the results of top-performing fully supervised video salient object detection methods.

In summary, this paper has the following contributions:

• We introduce a refinement network equipped with a non-locally enhanced recurrent module to generate saliency maps with spatiotemporal coherence.

• We further propose a flow-guided pseudo-label generator, which captures the interframe continuity of video and generates pseudo-labels of intervals based on sparse annotations.

• Under the joint supervision of the generated pseudo-labels and the manually labeled sparse annotations (e.g., 20% ground truth labels), our semi-supervised model can be trained to outperform existing state-of-the-art fully supervised video salient object detection methods.

2. Related Work

2.1. Salient Object Detection

Benefiting from the development of deep convolutional networks, salient object detection has recently achieved significant progress. In particular, these methods based on the fully convolutional network (FCN) and its variants [23, 12, 26] have become the dominant methods in this field, due to their powerful end-to-end feature learning nature and high computational efficiency. Nevertheless, these methods are inapplicable to video salient object detection without considering spatiotemporal information and contrast information within both motion and appearance in videos. Recently, attempts to apply deep CNNs to video salient object detection have attracted considerable research interest. Wang *et al.* [44] introduced FCN to this problem by taking adjacent pairs of frames as input. However, this method fails to learn sufficient spatiotemporal information with a limited number of input frames. To overcome this deficiency, Li *et al.* [24] proposed to enhance the temporal coherence at the feature level by exploiting both motion information and sequential feature evolution encoding. Fan *et al.* [10] proposed to captures video dynamics with a saliency-shift-aware module that learns human attention-shift. However, all the above methods rely on densely annotated video datasets, and none of them have ever attempted to reduce the dependence on dense labeling.

To the best of our knowledge, we are the first to explore the video salient object detection task by reducing the dependence on dense labeling. Moreover, we verify that the generated pseudo-labels can overcome the ambiguity in the labeling process to some extent, thus facilitating our model to achieve better performance.

2.2. Video Object Segmentation

Video object segmentation tasks can be divided into two categories, including semi-supervised video object segmentation [16, 7] and unsupervised video object segmentation [38, 17]. Semi-supervised video object segmentation aims at tracking a target mask given from the first annotated frame in the subsequent frames, while unsupervised video object segmentation aims at detecting the primary objects through the whole video sequence automatically.

It should be noted that the supervised or semi-supervised video segmentation methods mentioned here are all for the test phase, and the training process of both tasks is fully supervised. The semi-supervised video salient object detection considered in this paper is aimed at reducing the labeling dependence of training samples during the training process. Here, unsupervised video object segmentation is the most related task to ours as both tasks require no annotations during the inference phase. It can be achieved by graph cut [33], saliency detection [42], motion analysis [28], or object proposal ranking [21]. Recently, unsupervised video object segmentation methods have been mainly based on deep learning networks, such as two-stream architecture [17], FCN network [5], and recurrent networks [38]. However, most of the deep learning methods rely on a large quantity of pixel-wise labels for fully supervised training.

In this paper, we address the semi-supervised video salient object detection task using pseudo-labels with a few annotated frames. Although our proposed model is trained with semi-supervision, it is still well applicable to unsupervised video object segmentation.

3. Our Approach

In this section, we elaborate on the details of the proposed framework for semi-supervised video salient object detection, which consists of three major components. First, a residual connected refinement network is proposed to provide a spatial feature extractor and a pixel-wise classifier for salient object detection, which are respectively used for extracting spatial saliency features from raw input images and encoding the features to pixel-wise saliency maps with low-level cues connected to high-level features. Second, a non-locally enhanced recurrent module is designed to enhance the spatiotemporal coherence of the feature representation. Finally, a flow-guided pseudo-label generation (FGPLG) model, comprised of a modified RCRNet and an off-the-shelf FlowNet 2.0 model [13], is applied to generate in-between pseudo-labels from sparsely annotated video frames. With appropriate numbers of pseudo-labels, RCRNet with the NER module can be trained to capture the spatiotemporal information and generate accurate saliency maps for dense input frames.

3.1. Refinement Network with Residual Connection

Typical deep convolutional neural networks can extract high-level features from low-level cues of images, such as colors and textures, using a stack of convolutional layers and downsampling operations. The downsampling operation obtains an abstract feature representation by gradually increasing the receptive field of the convolutional layers. However, many spatial details are lost in this process. Without sufficient spatial details, pixel-wise prediction tasks, such as salient object detection, cannot precisely predict
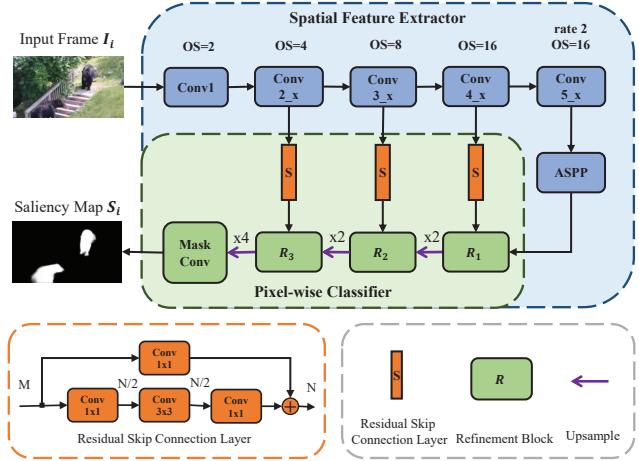

Figure 2. The architecture of our refinement network with residual connection (RCRNet). Here, '$\oplus$' denotes element-wise addition. Output stride (OS) explains the ratio of the input image size to the output feature map size.

on object boundaries or small objects. Inspired by [23], we adopt a refinement architecture to incorporate low-level spatial information in the decoding process for pixel-level saliency inference. As shown in Fig. 2, the proposed RCRNet consists of a spatial feature extractor $\mathcal{N}_{feat}$ and a pixel-wise classifier $\mathcal{N}_{seg}$ connected by three connection layers in different stages. The output saliency map S of a given frame I can be computed as

$$S = \mathcal{N}_{seg}(\mathcal{N}_{feat}(I)). \tag{1}$$

Spatial Feature Extractor: The spatial feature extractor is based on a ResNet-50 [11] model. Specifically, we use the first five groups of layers of ResNet-50 and remove the downsampling operations in conv5_x to reduce the loss of spatial information. To maintain the same receptive field, we use dilated convolutions [48] with $rate = 2$ to replace the convolutional layers in the last layer. Then we attach an atrous spatial pyramid pooling (ASPP) [4] module to the last layer, which captures both the image-level global context and the multiscale spatial context. Finally, the spatial feature extractor produces features with 256 channels and $1/16$ of the original input resolution (OS = 16).

Pixel-wise Classifier: The pixel-wise classifier is composed of three cascaded refinement blocks, each of which is connected to a layer in the spatial feature extractor via a connection layer. It is designed to mitigate the impact of the loss of spatial details during the downsampling process. Each refinement block takes as input the previous bottom-up output feature map and its corresponding feature map connected from the top-down stream. The resolution of these two feature maps should be consistent, so the upsampling operation is performed via bilinear interpolation when necessary. The refinement block works by first

Non-locally Enhanced Temporal Module

Spatial Feature Extractor

DB-ConvGRU

Pixel-wise Classifier

Non-local Block

Non-local Block

Input video clip $\{I_1, ..., I_T\}$
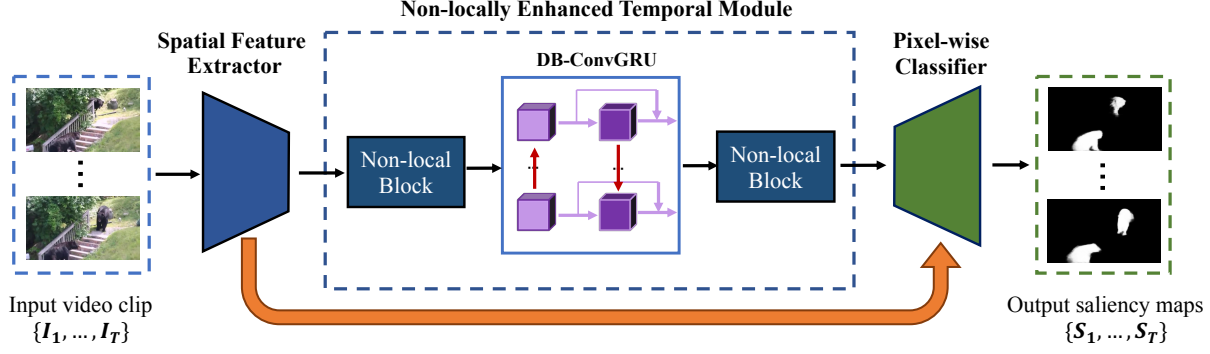
Output saliency maps $\{S_1, ..., S_T\}$

Figure 3. The architecture of our proposed video salient object detection network (RCRNet+NER). We incorporate a non-locally enhanced temporal module with our proposed RCRNet for spatiotemporal coherence modeling.

concatenating the feature maps and then feeding them to another 3×3 convolutional layer with 128 channels. Motivated by [11], a residual bottleneck architecture, named residual skip connection layer, is employed as the connection layer to connect low-level features to high-level ones. It downsamples the low-level feature maps from M channels to $N = 96$ channels and brings more spatial information to the refinement block. Residual learning allows us to connect the pixel-wise classifier to the pretrained spatial feature extractor without breaking its initial state (e.g., if the weight of the residual bottleneck is initialized as zero).

3.2. Non-locally Enhanced Recurrent Module

Given a sequence of video clip $I_i, i = 1, 2, ..., T$, video salient object detection aims at producing the saliency maps of all frames $S_i, i = 1, 2, ..., T$. Although the proposed RCRNet specializes in spatial saliency learning, it still lacks spatiotemporal modeling for video frames. Thus, we further propose a non-locally enhanced temporal (NER) module, which consists of two non-local blocks [45] and a convolutional GRU (ConvGRU) [1] module, to improve spatiotemporal coherence in high-level features. As shown in Fig. 3, incorporated with the NER module, RCRNet can be extended to video-based salient object detection.

Specifically, we first combine the features extracted from input video frames $\{I_i\}_{i=1}^{T}$ as $X = [X_1, X_2, ..., X_T]$. Here, $[,.,]$ denotes the concatenation operation and the spatial feature X_i of each frame I_i is computed as $X_i = \mathcal{N}_{feat}(I_i)$. Then, the combined feature X is fed into a non-local block. The non-local block computes the response at a position as a weighted sum of features at all positions for input feature maps. It can construct the spatiotemporal connection between the features of input video frames.

On the other hand, as a video sequence is composed of a series of scenes that are captured in chronological order, it is also necessary to characterize the sequential evolution of appearance contrast in the temporal domain. Based on this, we propose to exploit ConvGRU [1] modules for se-

quential feature evolution modeling. ConvGRU is an extension of traditional fully connected GRU [6] that has convolutional structures in both input-to-state and state-to-state connections. Let $\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_t$ denote the input to ConvGRU and $\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_t$ stand for its hidden states. A ConvGRU module consists of a reset gate $\mathcal{R}_t$ and an update gate $\mathcal{Z}_t$. With these two gates, ConvGRU can achieve selective memorization and forgetting. Given the above definition, the overall updating process of ConvGRU unrolled by time can be listed as follows:

$$\begin{aligned}
\mathcal{Z}_t &= \sigma(\mathcal{W}_{xz} * \mathcal{X}_t + \mathcal{W}_{hz} * \mathcal{H}_{t-1}), \\
\mathcal{R}_t &= \sigma(\mathcal{W}_{xr} * \mathcal{X}_t + \mathcal{W}_{hr} * \mathcal{H}_{t-1}), \\
\mathcal{H}'_t &= \tanh(\mathcal{W}_{xh} * \mathcal{X}_t + \mathcal{R}_t \circ (\mathcal{W}_{hh} * \mathcal{H}_{t-1})), \\
\mathcal{H}_t &= (1 - \mathcal{Z}_t) \circ \mathcal{H}'_t + \mathcal{Z}_t \circ \mathcal{H}_{t-1},
\end{aligned} \tag{2}$$

where '$*$' denotes the convolution operator and '$\circ$' denotes the Hadamard product. $\sigma(.)$ represents the sigmoid function and $\mathcal{W}$ represents the learnable weight matrices. For notational simplicity, the bias terms are omitted.

Motivated by [36], we stack two ConvGRU modules with forward and backward directions to strengthen the spatiotemporal information exchanges between two directions. In this way, deeper bidirectional ConvGRU (DB-ConvGRU) can memorize not only past sequences but also future ones. It can be formulated as follows:

$$\begin{aligned}
\mathcal{H}_t^f &= \text{ConvGRU}(\mathcal{H}_{t-1}^f, X_t), \\
\mathcal{H}_t^b &= \text{ConvGRU}(\mathcal{H}_{t+1}^b, \mathcal{H}_t^f), \\
\mathcal{H}_t &= \tanh(\mathcal{W}_{hf} * \mathcal{H}_t^f + \mathcal{W}_{hb} * \mathcal{H}_t^b),
\end{aligned} \tag{3}$$

where $\mathcal{H}_t^f$ and $\mathcal{H}_t^b$ represent the hidden state from forward and backward ConvGRU units, respectively. $\mathcal{H}_t$ represents the final output of DB-ConvGRU. X_t is the tth output feature from the non-local block.

As proven in [45], more non-local blocks in general lead to better results. Thus, we attach another non-local block to DB-ConvGRU to further enhance spatiotemporal coherence.
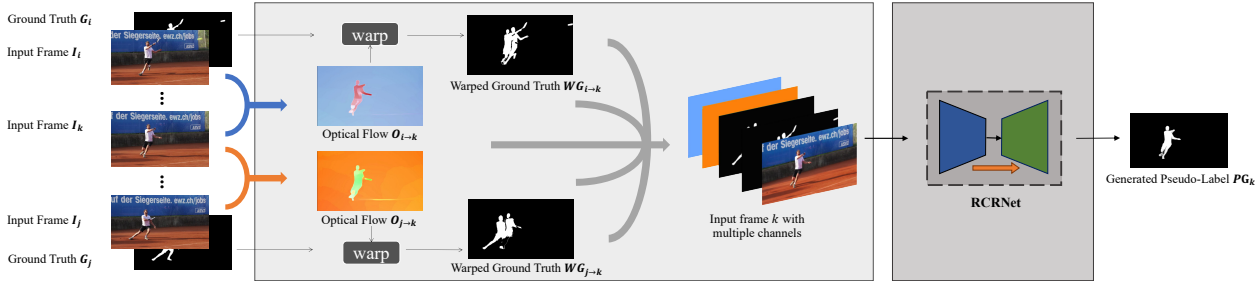
Figure 4. The architecture of our proposed flow-guided pseudo-label generation model (FGPLG).

3.3. Flow-Guided Pseudo-Label Generation Model

Although the proposed RCRNet+NER has a great potential to produce saliency maps with spatiotemporal coherence. With only a few sparsely annotated frames, it can barely learn enough temporal information, which greatly reduces the temporal coherence of the resulting saliency maps. To solve this problem, we attempt to generate denser pseudo-labels from a few sparse annotations and train our video saliency model with both types of labels.

Given triplets of input video frames $\{I_i, I_k, I_j\}(i < k < j)$, the proposed FGPLG model aims at generating a pseudo-label for frame I_k with ground truth G_i and G_j propagated from frame I_i and I_j, respectively. First, it computes the optical flow $O_{i \rightarrow k}$ from frame I_i to frame I_k with the off-the-shelf FlowNet 2.0. The optical flow $O_{j \rightarrow k}$ is obtained in the same way. Then, the label of frame I_k is estimated by applying a warping function to adjacent ground truth G_i and G_j. Nevertheless, as we can see in Fig. 4, the warped ground truth $WG_{i \rightarrow k}$ and $WG_{j \rightarrow k}$, are still too noisy to be used as supervisory information for practical training. Although the magnitude of optical flow $\|O_{i \rightarrow k}\|$ and $\|O_{j \rightarrow k}\|$ provide reasonable estimations of the motion mask of frame i_k, they cannot be employed as the estimated ground truth directly since not all the motion masks are salient. To further refine the estimated pseudo-label of frame I_k, another RCRNet is modified to accept a frame I_k^+ with 7 channels including RGB channels of frame I_k, adjacent warped ground truth $WG_{i \rightarrow k}$ and $WGj \rightarrow k$ and optical flow magnitude $\|O_{i \rightarrow k}\|$ and $\|O_{j \rightarrow k}\|$. With the above settings, a more reasonable and precise pseudo-label PG_k of frame I_k can be generated as:

$$PG_k = N_{seg}(N_{feat}(I_k^+)). \tag{4}$$

Here, the magnitude of optical flow is calculated by first normalizing the optical flow into interval $[-1, 1]$ and then computing its Euclidean norm.

The generation model can be trained with sparsely annotated frames to generate denser pseudo-labels. In our experiments, we use a fixed interval l to select sparse annotations for training. We take an annotation every l frames, i.e., the interval between the jth and k^{th} frame, and the interval be-

tween the ith and k^{th} frame are both equal to l. Experimental results show that the generation model designed in this way has a strong generalization ability. It can use the model trained by the triples sampled at larger interframe intervals to generate dense pseudo-labels of very high quality.

4. Experimental Results

4.1. Datasets and Evaluation

We evaluate the performance of our method on three public datasets: VOS [27], DAVIS [35] and FBMS [3]. VOS is a large-scale dataset with 200 indoor/outdoor videos for video-based salient object detection. It contains 116,103 frames including 7,650 pixel-wise annotated keyframes. The DAVIS dataset contains 50 high-quality videos, with a total of 3,455 pixel-wise annotated frames. The FBMS dataset contains 59 videos, totaling 720 sparsely annotated frames. We evaluate our trained RFCN+NER on the test sets of VOS, DAVIS, and FBMS for the task of video salient object detection.

We adopt precision-recall curves (PR), maximum F-measure and S-measure for evaluation. The F-measure is defined as $F_\beta = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}$. Here, β^2 is set to 0.3 as done by most existing image-based models [2, 23, 12]. We report the maximum F-measure computed from all precision-recall pairs. The S-measure is a new measure proposed in [9], which can simultaneously evaluate both region-aware and object-aware structural similarity between a saliency map and its corresponding ground truth.

4.2. Implementation Details

Our proposed method is implemented on PyTorch [34], a flexible open source deep learning platform. First, we initialize the weights of the spatial feature extractor in RCR-Net with an ImageNet [8] pretrained ResNet-50 [11]. Next, we pretrain the RCRNet using two image saliency datasets, i.e., MSRA-B [31] and HKU-IS [25], for spatial saliency learning. For semi-supervised video salient object detection, we combine the training sets of VOS [27], DAVIS [35], and FBMS [3] as our training set. The RCRNet pretrained on image saliency datasets is used as the backbone of the

Datasets	Metric	I+C			I+D							V+U			V+D			Ours*
		MC	RBD	MB+	RFCN	DCL	DHS	DSS	MSR	DGRL	PiCA	SAG	GF	SSA	FCNS	FGRN	PDB	
VOS	$F_\beta^{max} \uparrow$	0.558	0.589	0.577	0.680	0.704	0.715	0.703	0.719	0.723	0.734	0.541	0.529	0.669	0.681	0.714	0.741	0.856
	$S \uparrow$	0.612	0.652	0.638	0.721	0.728	0.783	0.760	0.764	0.776	0.796	0.597	0.560	0.710	0.727	0.734	0.797	0.872
DAVIS	$F_\beta^{max} \uparrow$	0.488	0.481	0.520	0.732	0.760	0.785	0.775	0.775	0.758	0.809	0.519	0.619	0.697	0.764	0.797	0.849	0.859
	$S \uparrow$	0.590	0.620	0.568	0.788	0.803	0.820	0.814	0.789	0.811	0.844	0.663	0.686	0.738	0.757	0.838	0.878	0.884
FBMS	$F_\beta^{max} \uparrow$	0.466	0.488	0.540	0.764	0.760	0.765	0.776	0.809	0.813	0.823	0.545	0.609	0.597	0.752	0.801	0.823	0.861
	$S \uparrow$	0.567	0.591	0.586	0.765	0.772	0.793	0.793	0.835	0.832	0.847	0.632	0.642	0.634	0.747	0.818	0.839	0.870

* Note that our model is a semi-supervised learning model using only approximately 20% ground truth labels for training.

Table 1. Comparison of quantitative results using maximum F-measure $F_\beta^{max} \uparrow$ (larger is better), S-measure $S \uparrow$ (larger is better). The best three results on each dataset are shown in **red**, **blue**, and **green**, respectively. Symbols of model categories: I+C for image-based classic unsupervised or non-deep learning methods, I+D for image-based deep learning methods, V+U for video-based unsupervised methods, V+D for video-based deep learning methods. Refer to the supplemental document for more detailed results.
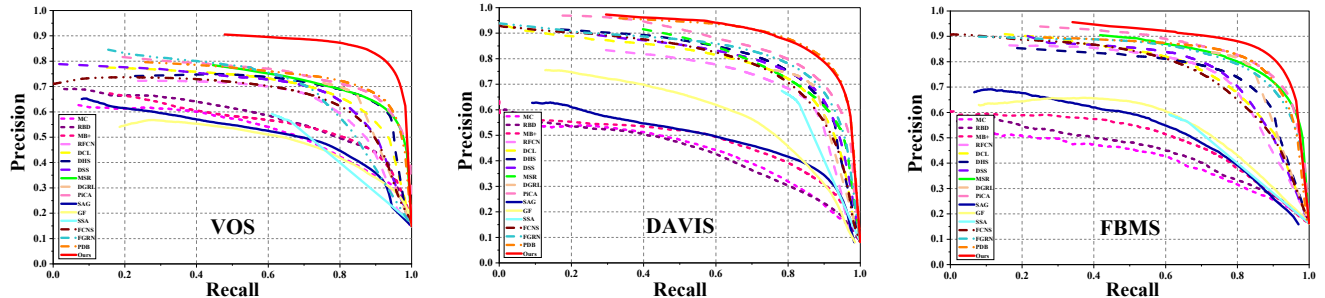

Figure 5. Comparison of precision-recall curves of 15 saliency detection methods on the VOS, DAVIS and FBMS datasets. Our proposed RCRNet+NER consistently outperforms other methods across three testing datasets using only 20% of ground truth labels.

pseudo-label generator. Then the FGPLG fine-tuned with a subset of the video training set is used to generate pseudo-labels. By utilizing the pseudo-labels together with the subset, we jointly train the RCRNet+NER, which takes a video clip of length T as input, to generate saliency maps to all input frames. Due to the limitation of machine memory, the default value of T is set to 4 in our experiments.

During the training process, we adopt Adam [19] as the optimizer. The learning rate is initially set to 1e-4 when training RCRNet, and is set to 1e-5 when fine-tuning RCRNet+NER and FGPLG. The input images or video frames are resized to 448×448 before being fed into the network in both training and inference phases. We use sigmoid cross-entropy loss as the loss function and compute the loss between each input image/frame and its corresponding label, even if it is a pseudo-label. In Section 4.4, we explore the effect of different amount of ground truth (GT) and pseudo-labels usage. It shows that when we take one GT and generate one pseudo-label every five frames (column '1 / 5' in Table 2) as the new training set, RCRNet+NER can be trained to outperform the model trained with all ground truth labels on the VOS dataset. We use this setting when performing external comparisons with existing state-of-the-art methods. In this setting, it takes approximately 10 hours to finish the whole training process on a workstation with an NVIDIA GTX 1080 GPU and a 2.4 GHz Intel CPU. In the inference phase, it takes approximately 37 ms to generate a saliency map for a 448×448 input frame, which reaches a real-time speed of 27 fps.

4.3. Comparison with State-of-the-Art

We compare our video saliency model (RCRNet+NER) against 16 state-of-the-art image/video saliency methods, including MC [18], RBD [50], MB+ [49], RFCN [39], DCL [26], DHS [29], DSS [12], MSR [23], DGRL [40], PiCA [30], SAG [42], GF [43], SSA [27], FCNS [44], FGRN [24], and PDB [36]. For a fair comparison, we use the implementations provided by the authors and fine-tune all the deep learning-based methods using the same training set, as mentioned in Section 4.2.

A visual comparison is given in Fig. 6. As shown in the figure, RCRNet+NER can not only accurately detect salient objects but also generate precise and consistent saliency maps in various challenging cases. As a part of the quantitative evaluation, we show a comparison of PR curves in Fig. 5. Moreover, a quantitative comparison of maximum F-measure and S-measure is listed in Table 1. As can be seen, our method can outperform all the state-of-the-art image-based and video-based saliency detection methods on VOS, DAVIS, and FBMS. Specifically, our RCRNet+NER improves the maximum F-measure achieved by the existing best-performing algorithms by 15.52%, 1.18%, and 4.62% respectively on VOS, DAVIS, and FBMS, and improves the S-measure by 9.41%, 0.68%, 2.72% accordingly. It is worth noting that our proposed method uses only approximately 20% ground truth maps in the training process to outperform the best-performing fully supervised video-based method (PDB), even though both models are based on the same backbone network (ResNet-50).
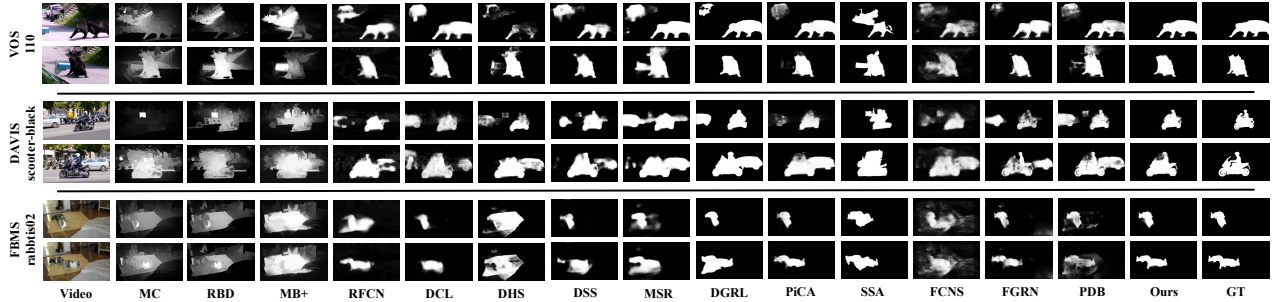
Figure 6. Visual comparison of saliency maps generated by state-of-the-art methods, including our RCRNet+NER. The ground truth (GT) is shown in the last column. Our model consistently produces saliency maps closest to the ground truth. Zoom in for details.
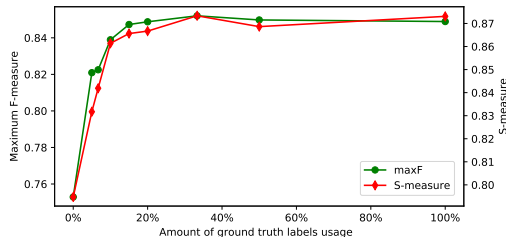

Figure 7. Sensitivities analysis on the amount of ground truth labels usage.

Labels	m/l	0 / 1	0 / 2	0 / 5	1 / 5	4 / 5	0 / 20	7 / 20	19 / 20
Proportion	GT	100%	50%	20%	20%	20%	5%	5%	5%
	Pseudo	0%	0%	0%	20%	80%	0%	35%	95%
Metric	$F_\beta^{max} \uparrow$	0.849	0.850	0.849	**0.861**	0.850	0.821	0.847	0.845
	$S \uparrow$	0.873	0.869	0.867	**0.874**	0.873	0.832	0.861	0.860

Table 2. Some representative quantitative results on different amount of ground truth (GT) and pseudo-labels usage. Here, l refers to the GT label interval, and m denotes the number of pseudo-labels used in each interval. For example, '0 / 5' means using one GT every five frames with no pseudo-labels. '1 / 5' means using one GT and generating one pseudo-label every five frames. Refer to the supplemental document for more detailed analysis.

4.4. Sensitivities to Different Amount of Ground Truth and Pseudo-Labels Usage

As described in Section 4.3, RCRNet+NER achieves state-of-the-art performance using only a few GTs and generated pseudo-labels for training. To demonstrate the effectiveness of our proposed semi-supervised framework, we explore the sensitivities to different amount of GT and pseudo-labels usage on the VOS dataset. First, we take a subset of the training set of VOS by a fixed interval and then fine-tune the RCRNet+NER with it. By repeating the above experiment with different fixed intervals, we show the performance of RCRNet+NER trained with different number of GT labels in Fig. 7. As shown in the figure, when the number of GT labels is severely insufficient (e.g., 5% of the origin training set), RCRNet+NER can benefit substantially from the increase in GT label usage. An interesting phenomenon is that when the training set is large enough, the application of denser label data does not necessarily lead to

better performance. Considering that adjacent densely annotated frames share small differences, ambiguity is usually inevitable during the manual labeling procedure, which may lead to overfitting and affect the generalization performance of the model.

Then, we further use the proposed FGPLG to generate different number of pseudo-labels with different number of GT labels. Some representative quantitative results are shown in Table 2, where we find that when there are insufficient GT labels, adding an appropriate number of generated pseudo-labels for training can effectively improve the performance. Furthermore, when we use 20% of annotations and 20% of pseudo-labels (column '1 / 5' in the table) to train RCRNet+NER, it reaches the max $F_\beta = 0.861$ and S-measure $= 0.874$ on the test set of VOS, surpassing the one trained with all GT labels. Even if trained with 5% of annotations and 35% of pseudo-labels (column '7 / 20' in the table), our model can produce comparable results. This interesting phenomenon demonstrates that pseudo-labels can overcome labeling ambiguity to some extent. Moreover, it also indicates that it is not necessary to densely annotate all video frames manually considering redundancies. Under the premise of the same labeling effort, selecting the sparse labeling strategy to cover more kinds of video content, and assisting with the generated pseudo-labels for training, will bring more performance gain.

4.5. Ablation Studies

To investigate the effectiveness of the proposed modules, we conduct the ablation studies on the VOS dataset.
The effectiveness of NER. As described in Section 3.2, our proposed NER module contains three cascaded modules, including a non-local block, a DB-ConvGRU module, and another non-local block. To validate the effectiveness and necessity of each submodule, we compare our RCRNet equipped with NER or its four variants on the test set of VOS. Here, we use one ground truth and one pseudo-label every five frames as the training set, to fix the impact of different amount of GT and pseudo-labels usage. As shown in Table 3, R_e refers to our proposed RCRNet with a non-locally enhanced module. By comparing the perfor-

mance of the first three variants R_a, R_b, and R_c, we find that adding non-local blocks and DB-ConvGRU can create a certain level of performance improvement. On the basis of R_c, adding an extra non-local block (i.e., R_e) can further increase 0.5% w.r.t max F-measure. When compared with R_d and R_e, we observe that DB-ConvGRU is indeed superior to ConvGRU as it involves deeper bidirectionally sequential modeling.

The effectiveness of FGPLG. As mentioned in Section 3.3, FGPLG model takes multiple channels as input to generate pseudo-labels, including image RGB channels, warped adjacent ground truth maps, and magnitude of optical flow. To validate the effectiveness and necessity of each component, we train three separate RCRNet+NER with pseudo-labels generated by our proposed FGPLG including its two variants, each of which takes different channels as input. Here, we use one ground truth and seven pseudo-labels every 20 frames as the training set for comparison. It also includes the performance of model G_a, which is trained without pseudo-labels, as a baseline. As shown in Table 4, the models trained with pseudo-labels (i.e., G_b, G_c, and G_d) all surpass the baseline model G_a, which further validates the effectiveness of using pseudo-labels for training. On the basis of G_b, adding adjacent ground truth as input (i.e., G_c) slightly improves the performance, while our proposed pseudo-label generator G_d outperforms all the other variants with a significant margin by further exploiting adjacent ground truth through flow-guided motion estimation.

Methods	R_a	R_b	R_c	R_d	R_e
ConvGRU?				✓	
DB-ConvGRU			✓		✓
first non-local block?		✓	✓	✓	✓
second non-local block?				✓	✓
$F_\beta^{max} \uparrow$	0.846	0.853	0.856	0.857	**0.861**
$S \uparrow$	0.865	0.871	0.871	0.872	**0.874**

Table 3. Effectiveness of non-locally enhanced recurrent module.

Methods	G_a	G_b	G_c	G_d
without label generation?	✓			
RGB channels?		✓	✓	✓
adjacent ground truth?			✓	✓
optical flow and GT warping?				✓
$F_\beta^{max} \uparrow$	0.821	0.832	0.838	**0.847**
$S \uparrow$	0.832	0.854	0.860	**0.861**

Table 4. Effectiveness of flow-guided label generation model.

Dataset	Metric	Methods						
		PDB	LVO	FSEG	LMP	SFL	FST	Ours*
DAVIS	$\mathcal{J} \uparrow$	74.3	70.1	70.7	70.0	67.4	55.8	**74.7**
	$\mathcal{F} \uparrow$	72.8	72.1	65.3	65.9	66.7	51.1	**73.3**
FBMS	$\mathcal{J} \uparrow$	72.3	65.1	68.4	35.7	35.7	47.7	**75.9**

* Note that our model is a semi-supervised learning model using only approximately 20% ground truth labels for training.

Table 5. Performance comparison with 6 representative unsupervised video object segmentation methods on the DAVIS and FBMS datasets. The best scores are marked in **bold**.

5. Performance on Unsupervised Video Object Segmentation

Unsupervised video object segmentation aims at automatically separating primary objects from input video sequences. As described, its problem setting is quite similar to video salient object detection, except that it seeks to perform a binary classification instead of computing a saliency probability for each pixel. To demonstrate the advantages and generalization ability of our proposed semi-supervised model, we test the pretrained RCRNet+NER (mentioned in Section 4) on the DAVIS and FBMS dataset without any pre-/post-processing and make a fair comparison with other 6 representative state-of-the-art unsupervised video segmentation methods, including FST [33], SFL [5], LMP [37], FSEG [17], LVO [38] and PDB [36]. We adopt the mean Jaccard index $\mathcal{J}$ (intersection-over-union) and mean contour accuracy $\mathcal{F}$ as metrics for quantitative comparison on the DAVIS dataset according to its settings. For the FBMS dataset, we employ the mean Jaccard index $\mathcal{J}$, as done by previous works [36, 24]. As shown in Table 5, our proposed method outperforms the above methods on both the DAVIS and FBMS datasets, which implies that our method has a strong ability to capture spatiotemporal information from video frames and is applicable to unsupervised video segmentation.

6. Conclusion

In this paper, we propose an accurate and cost-effective framework for video salient object detection. Our proposed RCRNet equipped with a non-locally enhanced recurrent module can learn to effectively capture spatiotemporal information with only a few ground truths and an appropriate number of pseudo-labels generated by our proposed flow-guided pseudo-label generation model. We believe this will bring insights to future work on the manual annotation for video segmentation tasks. Experimental results demonstrate that our proposed method can achieve state-of-the-art performance on video salient object detection and is also applicable to unsupervised video segmentation. In future work, we will further explore the impact of the use of keyframe selection instead of interval sampling of GT labels on the performance of the proposed method.

Acknowledgements

References

[1] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. In *Proceedings of the International Conference on Learning Representations*, 2016. 4

[2] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015. 5

[3] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *Proceedings of European Conference on Computer Vision*, pages 282–295, 2010. 2, 5

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[5] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 686–695, 2017. 3, 8

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, page 17241734, 2014. 4

[7] Prakash Chockalingam, Nalin Pradeep, and Stan Birchfield. Adaptive fragments-based tracking of non-rigid objects using level sets. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1530–1537, 2009. 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[9] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4548–4557, 2017. 5

[10] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 8554–8564, 2019. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3, 4, 5

[12] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017. 2, 5, 6

[13] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2462–2470, 2017. 2, 3

[14] Laurent Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, 13(10):1304–1318, 2004. 1

[15] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (11):1254–1259, 1998. 2

[16] Suyog Dutt Jain and Kristen Grauman. Supervoxel-consistent foreground propagation in video. In *Proceedings of European Conference on Computer Vision*, pages 656–671, 2014. 2

[17] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2126, 2017. 2, 3, 8

[18] Bowen Jiang, Lihe Zhang, Huchuan Lu, Chuan Yang, and Ming-Hsuan Yang. Saliency detection via absorbing markov chain. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1665–1672, 2013. 6

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015. 6

[20] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, 2017. 1

[21] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1995–2002, 2011. 3

[22] Guanbin Li, Yuan Xie, and Liang Lin. Weakly supervised salient object detection using image labels. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1, 2

[23] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2386–2395, 2017. 2, 3, 5, 6

[24] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3243–3252, 2018. 1, 2, 6, 8

[25] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 5455–5463, 2015. 5

[26] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 478–487, 2016. 2, 6

[27] Jia Li, Changqun Xia, and Xiaowu Chen. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE Transactions on Image Processing*, 27(1):349–364, 2018. 1, 2, 5, 6

[28] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C-C Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *Proceedings of*

European Conference on Computer Vision, pages 207–223, 2018. 3

[29] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 678–686, 2016. 6

[30] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3089–3098, 2018. 6

[31] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):353–367, 2011. 5

[32] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. In *Proceedings of the Tenth ACM International Conference on Multimedia*, pages 533–542. ACM, 2002. 1

[33] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1777–1784, 2013. 3, 8

[34] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[35] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016. 2, 5

[36] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of European Conference on Computer Vision*, pages 715–731, 2018. 1, 4, 6, 8

[37] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3386–3394, 2017. 8

[38] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4481–4490, 2017. 2, 3, 8

[39] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *Proceedings of European Conference on Computer Vision*, pages 825–841, 2016. 6

[40] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3127–3135, 2018. 6

[41] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of IEEE*

Conference on Computer Vision and Pattern Recognition, pages 4894–4903, 2018. 1

[42] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3395–3402, 2015. 1, 3, 6

[43] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196, 2015. 1, 6

[44] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2018. 1, 2, 6

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 4

[46] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2314–2320, 2017. 1

[47] Hefeng Wu, Guanbin Li, and Xiaonan Luo. Weighted attentional blocks for probabilistic object tracking. *The Visual Computer*, 30(2):229–243, 2014. 1

[48] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *Proceedings of the International Conference on Learning Representations*, 2016. 3

[49] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Minimum barrier salient object detection at 80 fps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1404–1412, 2015. 6

[50] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2814–2821, 2014. 6

Semi-Supervised Video Salient Object Detection Using Pseudo-Labels
(Supplemental Materials)

Pengxiang Yan[1]   Guanbin Li[1*]   Yuan Xie[1,2]
Zhen Li[3]   Chuan Wang[4]   Tianshui Chen[1,2]   Liang Lin[1,2]

[1]Sun Yat-sen University   [2]DarkMatter AI Research   [4]Megvii Technology
[3]Shenzhen Research Institute of Big Data, the Chinese University of Hong Kong (Shenzhen)

yanpx@mail2.sysu.edu.cn, liguanbin@mail.sysu.edu.cn, xiey39@mail2.sysu.edu.cn,
lizhen36@connect.hku.hk, wangchuan@megvii.com, tianshuichen@gmail.com, linliang@ieee.org

1. More Comparison with State-of-the-Art

We compare our video saliency model (RCRNet+NER) against 16 state-of-the-art image/video saliency methods, including MC [3], RBD [20], MB+ [19], RFCN [14], DCL [6], DHS [8], DSS [2], MSR [4], DGRL [15], PiCA [9], SAG [16], GF [17], SSA [7], FCNS [18], FGRN [5], and PDB [13]. A more detailed quantitative comparison of maximum F-measure, S-measure, weighted F-measure, and mean absolute error (MAE) on VOS [7], DAVIS [12] and FBMS [1] datasets is presented in Table 1. The weighted F-measure is proposed in [10] to mitigate the interpolation flaw, dependency flaw, and equal-importance flaw of traditional evaluation metrics. Here, we use the code provided by the authors with the default setting. MAE is defined as the average absolute difference between the binary ground truth and the saliency map at the pixel level [11]. As shown in Table 1, our method outperforms all existing salient object detection algorithms across all datasets. Specifically, our method improves the maximum F-measure achieved by the existing best-performing algorithms by 15.52%, 1.18%, and 4.62% respectively on VOS, DAVIS, and FBMS, and improves the S-measure by 9.41%, 0.68%, 2.72% accordingly. Moreover, our method improves the weighted F-measure by 17.04%, 3.23%, and 3.68% respectively on VOS, DAVIS, and FBMS, and reduces the MAE by 34.67%, 6.67%, and 5.26% accordingly.

2. Sensitivities to Different Amount of Ground Truth and Pseudo-Labels Usage

To demonstrate the effectiveness of our proposed semi-supervised framework, we explore the sensitivities to different amount of ground truth and pseudo-labels usage on the VOS [7] dataset. We fine-tune our proposed video saliency

detector RCRNet+NER with different number of GT and pseudo-labels. Detailed quantitative results on the test set of VOS are presented in Table 2. As seen, models cannot generate temporally consistent saliency maps when the training data set is seriously deficient (e.g., 5%) , which results in inferior performance. Nevertheless, an interesting phenomenon is that when there are enough training data with similar appearance, given more annotation data does not guarantee continuous performance improvement. This phenomenon may be due to model overfitting caused by label ambiguity. Based on the above observations, we propose jointly training RCRNet+NER with an appropriate number of pseudo-labels (e.g., 20%) and GT labels (e.g., 20%). Experimental results demonstrate the effectiveness of using pseudo-labels for training. Moreover, our semi-supervised RCRNet+NER (column '1 / 5' in the table) can even outperform the one trained with all annotated frames.

References

[1] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *Proceedings of European Conference on Computer Vision*, pages 282–295, 2010. 1, 2

[2] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017. 1, 2

[3] Bowen Jiang, Lihe Zhang, Huchuan Lu, Chuan Yang, and Ming-Hsuan Yang. Saliency detection via absorbing markov chain. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1665–1672, 2013. 1, 2

[4] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2386–2395, 2017. 1, 2

---

*Corresponding author is Guanbin Li.

Table 1:

	Methods	Pub.	VOS [7] $F_\beta^{max}\uparrow$	$S\uparrow$	$F_\beta^w\uparrow$	MAE$\downarrow$	DAVIS [12] $F_\beta^{max}\uparrow$	$S\uparrow$	$F_\beta^w\uparrow$	MAE$\downarrow$	FBMS [1] $F_\beta^{max}\uparrow$	$S\uparrow$	$F_\beta^w\uparrow$	MAE$\downarrow$
I+C	MC [3]	CVPR'13	0.558	0.612	0.306	0.199	0.488	0.590	0.201	0.182	0.466	0.567	0.251	0.237
	RBD [20]	CVPR'14	0.589	0.652	0.419	0.148	0.481	0.620	0.264	0.142	0.488	0.591	0.307	0.194
	MB+ [19]	ICCV'15	0.577	0.638	0.401	0.167	0.520	0.568	0.231	0.229	0.540	0.586	0.319	0.244
I+D	RFCN [14]	ECCV'16	0.680	0.721	0.499	0.108	0.732	0.788	0.514	0.069	0.764	0.765	0.563	0.108
	DCL [6]	CVPR'16	0.704	0.728	0.551	0.086	0.760	0.803	0.578	0.065	0.760	0.772	0.630	0.088
	DHS [8]	CVPR'16	0.715	0.783	0.611	0.075	0.785	0.820	0.662	0.040	0.765	0.793	0.671	0.083
	DSS [2]	CVPR'17	0.703	0.760	0.587	0.082	0.775	0.814	0.621	0.056	0.776	0.793	0.656	0.083
	MSR [4]	CVPR'17	0.719	0.764	0.599	0.102	0.775	0.789	0.589	0.065	0.809	0.835	0.709	0.067
	DGRL [15]	CVPR'18	0.723	0.776	0.622	0.079	0.758	0.811	0.652	0.058	0.813	0.834	0.761	0.057
	PiCA [9]	CVPR'18	0.734	0.796	0.633	0.090	0.809	0.844	0.660	0.044	0.823	0.847	0.730	0.058
V+U	SAG [16]	CVPR'15	0.541	0.597	0.309	0.178	0.519	0.663	0.329	0.107	0.545	0.632	0.345	0.176
	GF [17]	TIP'15	0.529	0.560	0.318	0.264	0.619	0.686	0.374	0.094	0.609	0.642	0.365	0.172
	SSA [7]	TIP'18	0.669	0.710	0.580	0.117	0.697	0.738	0.601	0.068	0.597	0.634	0.496	0.165
V+D	FCNS [18]	TIP'18	0.681	0.727	0.439	0.125	0.764	0.757	0.407	0.091	0.752	0.747	0.475	0.129
	FGRN [5]	CVPR'18	0.714	0.734	0.547	0.082	0.797	0.838	0.662	0.044	0.801	0.818	0.643	0.083
	PDB [13]	ECCV'18	0.741	0.797	0.632	0.080	0.849	0.878	0.744	0.030	0.823	0.839	0.732	0.067
	Ours*	ICCV'19	0.856	0.872	0.776	0.049	0.859	0.884	0.768	0.028	0.861	0.870	0.789	0.054

* Note that our model is a semi-supervised learning model using only about 20% ground truth labels for training.

Table 1. Comparison of quantitative results using maximum F-measure F_β^{max} (larger is better), S-measure S (larger is better), weighted F-measure F_β^w (larger is better), and MAE (smaller is better). The best three results on each dataset are shown in **red**, **blue**, and **green**, respectively. Symbols of model categories: I+C for image-based classic unsupervised or non-deep learning methods, I+D for image-based deep learning methods, V+U for video-based unsupervised methods, V+D for video-based deep learning methods.

Labels	m / l	0 / 1	0 / 2	0 / 3	2 / 3	0 / 5	1 / 5	2 / 5	4 / 5	0 / 7	1 / 7	2 / 7	6 / 7
Proportion	GT	100%	50%	33.3%	33.3%	20%	20%	20%	20%	14.3%	14.3%	14.3%	14.3%
	Pseudo	0%	0%	0%	0%	0%	20%	40%	80%	0%	14.3%	28.6%	85.7%
Metric	$F_\beta^{max}\uparrow$	0.849	0.850	0.852	0.851	0.849	**0.861**	0.851	0.850	0.847	0.856	0.851	0.846
	$S\uparrow$	0.873	0.869	0.873	0.873	0.867	**0.874**	0.869	0.873	0.866	0.872	0.869	0.867

Labels	m / l	0 / 10	2 / 10	3 / 10	9 / 10	0 / 15	4 / 15	5 / 15	14 / 15	0 / 20	6 / 20	7 / 20	19 / 20
Proportion	GT	10%	10%	10%	10%	6.7%	6.7%	6.7%	6.7%	5%	5%	5%	5%
	Pseudo	0%	20%	30%	90%	0%	26.7%	33.3%	93.3%	0%	30%	35%	95%
Metric	$F_\beta^{max}\uparrow$	0.839	0.859	0.850	0.849	0.823	0.849	0.851	0.849	0.821	0.847	0.847	0.845
	$S\uparrow$	0.861	0.872	0.868	0.867	0.842	0.866	0.868	0.865	0.832	0.861	0.861	0.860

Table 2. More quantitative results on different amount of ground truth (GT) and pseudo-labels usage. Here, l refers to GT label interval, and m denotes the number of pseudo-labels used in each interval. For example, '0 / 5' means using one GT every five frames with no pseudo-labels. '1 / 5' means using one GT and generating one pseudo-label every five frames. And so on.

[5] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3243–3252, 2018. 1, 2

[6] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 478–487, 2016. 1, 2

[7] Jia Li, Changqun Xia, and Xiaowu Chen. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE Transactions on Image Processing*, 27(1):349–364, 2018. 1, 2

[8] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 678–686, 2016. 1, 2

[9] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detec-

tion. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3089–3098, 2018. 1, 2

[10] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2014. 1

[11] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–740, 2012. 1

[12] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016. 1, 2

[13] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of European Conference on Computer Vision*, pages 715–731, 2018. 1, 2

[14] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *Proceedings of European Conference on Computer Vision*, pages 825–841, 2016. 1, 2

[15] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3127–3135, 2018. 1, 2

[16] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3395–3402, 2015. 1, 2

[17] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196, 2015. 1, 2

[18] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2018. 1, 2

[19] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Minimum barrier salient object detection at 80 fps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1404–1412, 2015. 1, 2

[20] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2814–2821, 2014. 1, 2