Cross-Frame Feature-Saliency Mutual Reinforcing for Weakly Supervised Video Salient Object Detection

Jian Wang[a,b,1], Siyue Yu[a,1], Bingfeng Zhang[c,1], Xinqiao Zhao[a], Ángel F. García-Fernández[b,d], Eng Gee Lim[a,*], Jimin Xiao[a,*]

[a] *Xi'an Jiaotong-Liverpool University, Suzhou, China*
[b] *University of Liverpool, Liverpool, UK*
[c] *China University of Petroleum (East China), Qingdao, China*
[d] *ARIES Research Centre, Universidad Antonio de Nebrija, Spain*

Abstract

Scribble annotations have recently become popular in video salient object detection. Previous methods only focus on utilizing shallow feature consistency for more integral predictions. However, there is potential for consistency between cross-frame deep features to be used to help regularize saliency predictions better. Besides, we have observed that leveraging saliency predictions as pseudo-supervision signals yields notable improvements in extracting both intra-frame and cross-frame deep features. This, in turn, leads to more precise and detailed object structural information. Thus, we propose a cross-frame feature-saliency mutual reinforcing training process to assist scribble annotations for integral video saliency predictions. Specifically, we design a cross-frame feature regularization head, which leverages intra-frame and cross-frame deep feature consistency to regularize saliency predictions as auxiliary supervision. Then, to help obtain more accurate feature consistency, we design a cross-frame saliency regularization head, where predicted saliency values are used as pseudo-supervision signals to acquire better feature consistency. In this way, our cross-frame fea-

*Corresponding authors
Email addresses: Jian.Wang21@student.xjtlu.edu.cn (Jian Wang), siyue.yu02@xjtlu.edu.cn (Siyue Yu), Bingfeng.Zhang@upc.edu.cn (Bingfeng Zhang), Xinqiao.Zhao20@student.xjtlu.edu.cn (Xinqiao Zhao), angel.garcia-fernandez@liverpool.ac.uk (Ángel F. García-Fernández), enggee.lim@xjtlu.edu.cn (Eng Gee Lim), jimin.xiao@xjtlu.edu.cn (Jimin Xiao)
[1]Co-first author. All the authors contributed equally to this research.

ture and saliency regularization heads can benefit from each other to help the network learn more accurately. Extensive experiments show that our method can achieve better performances than the previous best methods. Our source code will be publicly released.

Keywords: Video salient object detection, Scribble annotations, Cross-frame feature Consistency, Cross-frame Saliency Consistency

1. Introduction

Video salient object detection (VSOD) aims to detect and segment the most attractive objects in a video sequence. Different from salient object detection (SOD), which detects the salient objects in static images [1, 2, 3, 4, 5, 6, 7, 8, 9], VSOD needs to track the detected salient objects through all the frames. With the advances in neural networks, many fully supervised VSOD models [10, 11, 12, 13, 14] have achieved impressive performances and widely applied to different video processing tasks, such as video compression [15, 16], video object segmentation [17, 18, 19, 20], and video captioning [21, 22] by providing foreground and background indications. However, pixel-level annotations are required in such models, which are time-consuming and labor-intensive.

To reduce massive resources for pixel-level annotations, scribble annotations [23, 24, 25, 26, 27, 28] are gaining popularity due to their flexibility and efficiency. However, it is difficult for scribble annotations to provide integral object structure information since they only cover a small portion of the object. To get more accurate object structure and boundaries, previous methods [29, 30, 31] explore various target clues. For example, Zhao *et al.* [29] design an appearance-motion fusion module to acquire enhanced fused features from frames and optical flows to mine comprehensive target features. Besides, Gao *et al.* [30, 31] use shallow features (*e.g.*, RGB and position information) to propagate the annotations to unlabeled regions for integral target predictions. However, current methods only leverage shallow feature consistency on predictions as regularization, as illustrated in Fig. 1 (a). The fact that the con-

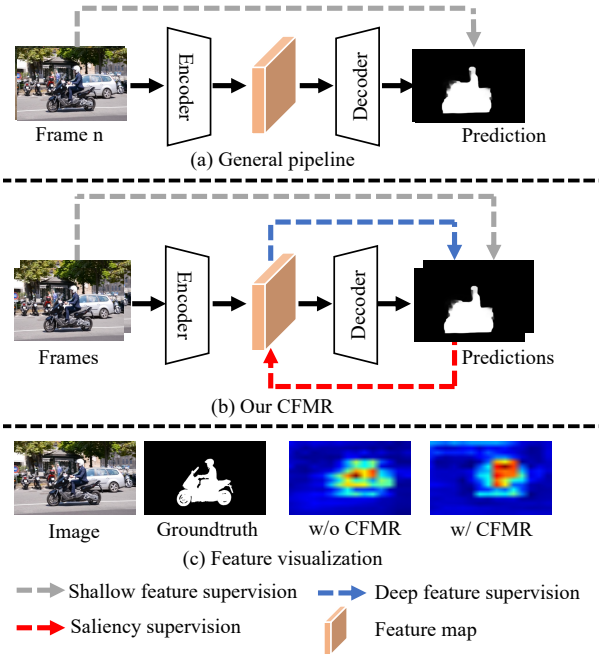

Figure 1: (a) General pipeline only uses intra-frame shallow features as auxiliary supervision. (b) Our cross-frame feature-saliency mutual reinforcing (CFMR) training process uses deep features (both intra-frame and cross-frame features), saliency predictions, and shallow features as supervision. (c) Feature visualization.

sistency between deep features (both intra-frame and cross-frame features) can help regularize saliency predictions has not been explored yet.

Although we find that the deep features possess much semantic information, some noise still exists, as shown in the third column of Fig. 1 (c). If we can introduce new supervision on the deep feature consistency of the same class pixels, the noise can be reduced, and the performance can be further enhanced. Considering this, we hypothesize that the saliency predictions can serve as valuable pseudo-supervision signals to assist the network in learning feature consistency relationships since they can provide more target object structural information than scribbles. Thus, we argue that better saliency predictions can help promote better deep feature consistency. Meanwhile, better feature consistency can help produce more integral saliency predictions.

To achieve this goal on weakly supervised VSOD (WSVSOD) with scribble annotations, we propose a **C**ross-frame **F**eature-saliency **M**utual **R**einforcing (CFMR) training process, where deep features and saliency predictions are used to supervise each other as demonstrated in Fig. 1 (b). Specifically, we design a cross-frame feature regularization head and a cross-frame saliency regularization head to realize this mutual reinforcing process. In our cross-frame feature regularization head, a cross-frame feature regularization (CFR) loss is designed to supervise saliency predictions, where both intra-frame and cross-frame deep feature consistency are utilized as criteria. On the other hand, our cross-frame saliency regularization head devotes to acquiring better feature representations. Specifically, we design a cross-frame saliency regularization (CSR) loss to make deep features belonging to the same salient object close to each other. Thus, the predicted saliency map is deployed to supervise deep feature consistency in turn. Similar to our CFR loss, we also involve cross-frame information as assistance to mine temporal context for more regularization in our CSR loss. As illustrated in Fig. 1 (c) ('w/ CFMR'), with the help of our CFMR, the background noise can be suppressed, and more relative target features can be aggregated. In this way, the mutual interaction of our cross-frame feature regularization head and cross-frame saliency regularization head can reinforce the network to learn comprehensive object structure information in a more accurate direction.

Overall, our contributions can be summarized as follows:

- We propose a Cross-frame Feature-saliency Mutual Reinforcing (CFMR) training process, including a cross-frame feature regularization head and a cross-frame saliency regularization head, to help learn comprehensive object structure information for scribble-supervised video salient object detection.

- We propose a cross-frame feature regularization (CFR) loss, where both intra-frame and cross-frame deep feature consistency are deployed to help supervise saliency predictions. Besides, we design a cross-frame saliency regularization (CSR) loss to help build a more accurate deep feature con-

sistency relationship. With the mutual interaction of our CFR loss and CSR loss, the network can learn more precisely for better saliency predictions.

- Our approach achieves new state-of-the-art performance, outperforming previous works on six widely used benchmarks. For instance, on SegV2 dataset [32], a gain of 2.7% for structure-based metric (S_α), 4.8% for maximum F-measure (F_β) and 0.8% for Mean Absolute Error (MAE) is obtained.

2. Related work

2.1. Fully-supervised video salient object detection

The last decade has witnessed significant improvements for fully supervised VSOD methods [33, 12, 10, 34]. According to how to explore temporal information, they can be divided into two types: convLSTM-based methods [12, 35, 29, 36, 37] and optical-flow-based methods [10, 11, 17, 38]. Wang *et al.* [39] are the first to propose a deep learning model for VSOD. Song *et al.* [36] propose a dilated bidirectional convLSTM to learn spatio-temporal information for long sequences. Fan *et al.* [12] develop a saliency shift-aware convLSTM to acquire temporal information. Moreover, Wang *et al.* [37] augment the CNN-LSTM architecture with a supervised attention mechanism to enable fast end-to-end saliency learning. Then, for optical-flow-based methods, Li *et al.* [10] design the motion saliency sub-network to aggregate optical flow with appearance information. In contrast to existing methodologies, CoSTFormer [40] integrates long-local and short-global spatial-temporal contexts through the utilization of two complementary transformer branches. A combination of these two branches enables explicit modeling of spatial-temporal relationships across various frames. Although such approaches achieve remarkable success in VSOD, they heavily rely on fully pixel-level annotations, which are time-consuming and notoriously expensive. To alleviate the high reliance on per-pixel labeling, we focus on WSVSOD with scribble annotations in this paper.

2.2. Weakly-supervised video salient object detection

In recent years, some works [29, 30] have explored deep learning methods in WSVSOD. Zhao *et al.* [29] come up with two re-labeled scribble datasets (DAVIS-S and DAVSOD-S) based on fixation-guided scribble annotations and design an appearance-motion fusion module to acquire enhanced fused features from frames and optical flows. Ma *et al.* [41] propose a novel two-step strategy for their model. In the initial step, they employ diverse strategies, including the Itti model [42], CAM [43], super-pixel [44] and DenseCRF [45] to generate four distinct pseudo-labels for each frame across all video datasets. Subsequently, they introduce various loss functions, namely edge loss, pseudo-label loss, self-supervised loss, and fusion loss, to effectively train their model using the pseudo-labels generated in the previous step. It can be inferred from the aforementioned pipeline that their focus primarily lies on individual frames within a video sequence while neglecting the cross-frame connections. In contrast, our approach incorporates two cross-frame heads that consider both intra-frame and cross-frame pixel relationships. Gao *et al.* [30] use a transformer-based module and shallow features (*e.g.*, RGB and position information) to propagate the scribble labels to unlabeled regions for integral target predictions. However, only shallow feature consistency is leveraged as regularization on saliency predictions. We argue that the consistency between deep features can also be used to help mine more comprehensive target predictions.

2.3. Weakly-supervised salient object detection

To alleviate the cost of labeling, many works [46, 31, 47] use weakly supervised or unsupervised approaches to detect salient objects. Zhang *et al.* [46] present the first weakly supervised SOD method where an auxiliary edge detection network is designed to enforce boundaries of predicted saliency. Yu *et al.* [31] design an AGGM module and a self-consistent mechanism to regularize the saliency map of different scales. Piao *et al.* [44] use class activation scores from class activation maps (CAMs) as clues to avoid the negative impacts of a single label. Gao *et al.* [48] use a transformer-based model and propose a

non-salient object suppression technique to explicitly filter out the non-salient objects with point annotation as supervision. Moreover, Li *et al.* [49] introduce a mutual information optimization method to explicitly model the contribution of RGB and depth for weakly-supervised RGB-D saliency detection. However, such methods cannot be directly used in VSOD due to the lack of temporal information. In this paper, scribble-supervised VSOD is studied.

3. Methodology

3.1. Overview

For simplicity, we explain our method through one video sequence in the training dataset, which is denoted as $\mathbf{I} = \{I^n\}_{n=1}^{N}$, where I^n is the sequence frame and N is the total number of the sequence. Scribble annotations are only provided in the training dataset. The overall framework of our approach is illustrated in Fig. 2. We first generate a pseudo current frame $\widetilde{I}^n$ by combining the previous frame I^{n-1} and the optical flow OF^{n-1}. Then, the pseudo current frame $\widetilde{I}^n$ and current frame I^n are input to the encoder for deep feature maps $\widetilde{f}^n$ and f^n respectively. Next, the deep features are transmitted to the video saliency network [29] to predict the corresponding video saliency maps $\widetilde{y}^n$ and y^n. After that, the pseudo current frame $\widetilde{I}^n$, current frame I^n, the deep features ($\widetilde{f}^n$ and f^n) and predictions (y^{n-1} and y^n) are sent to our cross-frame feature regularization head (Sect. 3.2.) and cross-frame saliency regularization head (Sect. 3.3) to conduct our CFMR training process. Meanwhile, the cross-frame saliency consistency (CSC) loss (Sect. 3.4), which optimizes the consistency between the predicted $\widetilde{y}^n$ and y^n, is applied as assistant supervision.

3.2. Cross-frame Feature Regularization Head

As mentioned before, it is difficult for VSOD networks to learn the object structure information solely relying on scribble annotations. Thus, our cross-frame feature regularization head aims to realize the feature consistency regularization on saliency predictions as auxiliary supervision. With this regularization, the scribble supervision information can be expanded to unlabeled

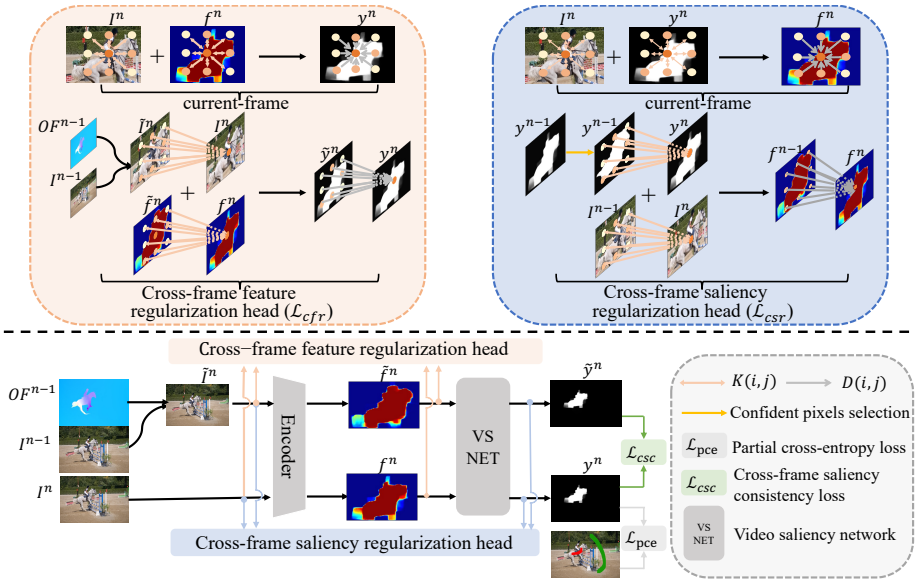

Figure 2: Our cross-frame feature-saliency mutual reinforcing (CFMR) training process and two regularization heads. First, a pseudo current frame $\widetilde{I}^n$ is generated using the previous frame I^{n-1} and the optical flow OF^{n-1}. Then, the pseudo current $\widetilde{I}^n$, current frame I^n are input to an encoder for deep semantic features $\widetilde{f}^n$ and f^n. The features are transmitted to the video saliency network (VSNET) to predict the corresponding saliency prediction $\widetilde{y}^n$ and y^n. Our cross-frame feature regularization head uses intra-frame and cross-frame feature consistency to help supervise saliency predictions. Our cross-frame saliency regularization head uses image information and saliency maps to supervise feature maps in turn. Additionally, $\mathcal{L}_{csc}$ is deployed to assist the two heads.

pixels to involve more pixels for training. Specifically, we design a CFR loss in our cross-frame feature regularization head based on the assumption that pixels of the current frame sharing similar features should have similar saliency values, and if pixels from different frames share similar features, they should also have similar saliency values. In this way, both spatial and temporal contexts can be harvested for abundant feature consistency relationships. However, it is difficult to obtain position relationships across different frames directly, we consider projecting the previous frame to the current frame with the help of optical flow:

$$\widetilde{I}^n = \mathcal{P}(I^{n-1}, OF^{n-1}), \tag{1}$$

where $\mathcal{P}(\cdot)$ means the mapping process and we utilize the model from RAFT [50] to achieve this mapping process. RAFT [50] can estimate the optical flow between video frames by creating multi-scale 4D correlation volumes for all pairs of pixels, and iteratively updating a flow field through a recurrent unit that performs lookups on the correlation volumes. By leveraging the excellent optical flow estimation capability of RAFT [50], a high quality pseudo current frame can be generated. Then, the generated pseudo current frame $\widetilde{I}^n$, which contains the original feature information of the previous frame I^{n-1}, is used for the cross-frame consistency in the following operations. With this mapping process, the moving position relationship can be well utilized for temporal information, and the feature consistency regularization between adjacent frames can be easily obtained.

We follow previous work [31] to define the pair-wise saliency similarity as:

$$D_y(y_i, y_j) = \|y_i - y_j\|, \tag{2}$$

where $\|\cdot\|$ is the L1 distance, y_i and y_j are the saliency prediction of pixel i and j in the current frame.

For computation efficiency, instead of computing the pair-wise relationship between all pixel pairs, we choose to compute the similarity of a reference point i with its adjacent points in a $k \times k$ kernel. For simplicity, we remove the superscript n in the following equations. Then, our cross-frame feature regularization loss ($\mathcal{L}_{cfr}$) is in the form of:

$$\begin{aligned}\mathcal{L}_{cfr} = &\underbrace{\frac{1}{M}\sum_{i=1}^{M}\sum_{j\in W_i}\left(K_1\left(i,j\right)D_y(y_i,y_j)\right)}_{current-frame} \\ &+\underbrace{\frac{1}{M}\sum_{i=1}^{M}\sum_{j\in \widetilde{W}_i}\left(K_2\left(i,j\right)D_y(y_i,\widetilde{y}_j)\right)}_{cross-frame},\end{aligned} \tag{3}$$

where W_i means the region belonging to a $k \times k$ kernel around pixel i. $\widetilde{W}_i$ means the same position region in the pseudo current frame. M means the

total number of pixels of the frame. $\widetilde{y}_j$ is the saliency prediction of pixel j in the pseudo current frame. $K_1(\cdot)$ and $K_2(\cdot)$ are the corresponding current-frame feature consistency and cross-frame feature consistency functions.

Considering the network cannot provide accurate saliency predictions initially, we also use shallow information (*i.e.*, RGB and position information) to help establish the feature consistency to supervise the saliency consistency. Therefore, the current feature consistency function follows the definition:

$$K_1(i,j) = e^{-\frac{\|p_i - p_j\|^2}{2\delta_p^2}} \cdot e^{-\frac{\|r_i - r_j\|^2}{2\delta_r^2}} \cdot e^{-\frac{\|f_i - f_j\|^2}{2\delta_f^2}}, \tag{4}$$

where $\|\cdot\|^2$ denotes the squared L2 distance, p_i and p_j denote the position of pixel i and pixel j. Variables r_i and r_j denote the RGB information of pixel i and j, and f_i and f_j denote the deep semantic features of pixel i and j. Variables $\{\delta_p, \delta_r, \delta_f\}$ are the hyper-parameters for the scale of the three pairwise similarities and are set to $\{6, 0.1, 50\}$. Note that the features are regarded as non-gradient values here. In this way, Eq. (4) can help leverage shallow feature consistency and deep feature consistency within the current frame.

Similarly, the cross-frame feature consistency function is defined as:

$$K_2(i,j) = e^{-\frac{\|p_i - \widetilde{p}_j\|^2}{2\delta_p^2}} \cdot e^{-\frac{\|r_i - \widetilde{r}_j\|^2}{2\delta_r^2}} \cdot e^{-\frac{\|f_i - \widetilde{f}_j\|^2}{2\delta_f^2}}, \tag{5}$$

where $\widetilde{p}_j$, $\widetilde{r}_j$ and $\widetilde{f}_j$ are position, RGB information and feature for pixel j in the generated pseudo current frame. In Eq. (5), by choosing to compute the pairwise similarity between pixel i and pixels around the same position of pixel i but in the pseudo current frame, the temporal information can be taken advantage of for more comprehensive feature consistency.

Our cross-frame feature regularization loss ($\mathcal{L}_{cfr}$) enforces similar pixels in the same kernel region to share consistent saliency scores. It can further assist scribble annotations by propagating labeled points to the whole frame during training and help the network learn integral and smooth salient regions with limited labels. Moreover, with the auxiliary of both the actual current frame and the pseudo current frame, our $\mathcal{L}_{cfr}$ can help the network learn in both spatial and temporal perspectives.

3.3. Cross-frame Saliency Regularization Head

Our cross-frame saliency regularization head is designed to improve the network's ability to learn precise feature consistency, as the third column of Fig. 1 (c) shows that much noise still persists in the deep features. To suppress the noise, we design a cross-frame saliency regularization loss using saliency predictions as pseudo-supervision signals to help CFR loss with correct feature consistency. Similar to our CFR loss, we argue that pixels from the same or different frames sharing similar saliency values tend to have similar features. We define the pair-wise feature consistency as:

$$D_f(f_i, f_j) = \|f_i - f_j\|, \tag{6}$$

where f_i and f_j are the feature vectors of pixel i and pixel j in the current frame.

We experimentally find that trivially using all the predicted saliency values from adjacent frames will cause performance degeneration. Thus, we only collect pixels with confident saliency values both in the foreground and background to help provide the cross-frame saliency consistency as follows:

$$B_f^n = \{\text{pixel } i |\ i \in I^n, y_i > \tau_f\}, \tag{7}$$

$$B_g^n = \{\text{pixel } i |\ i \in I^n, y_i < \tau_g\}, \tag{8}$$

where B_f^n is the set of pixels with confident saliency values in the foreground and B_g^n is the set of pixels with low saliency values in the background. τ_f and τ_g are hyper-parameters.

Selected pixel set B^n is the union of sets B_f^n and B_g^n.

$$B^n = B_f^n \bigcup B_g^n, \tag{9}$$

Then, similar to CFR loss, our cross-frame saliency regularization loss ($\mathcal{L}_{csr}$)

for a reference pixel i with a region of $k \times k$ kernel can be defined as:

$$\begin{aligned}\mathcal{L}_{csr} = & \underbrace{\frac{1}{M}\sum_{i=1}^{M}\sum_{j\in W_i}\left(K_1'\left(i,j\right)D_f(f_i^n, f_j^n)\right)}_{current-frame} \\ & + \underbrace{\frac{1}{|B^n|}\sum_{i\in B^n}\sum_{j\in \widehat{W}_i'}\left(K_2'\left(i,j\right)D_f(f_i^n, f_j^{n-1})\right)}_{cross-frame},\end{aligned} \tag{10}$$

where $K_1'(\cdot)$ and $K_2'(\cdot)$ are the corresponding current-frame saliency consistency and cross-frame saliency consistency functions. Here $\widehat{W}_i'$ means the pixels that satisfy the following condition:

$$\widehat{W}_i' = \widehat{W}_i \bigcap B^{n-1}, \tag{11}$$

where $\widehat{W}_i$ means the region with the same position as W_i in the previous frame.

The current-frame saliency consistency function is of the form:

$$K_1'(i,j) = e^{-\frac{\|p_i^n - p_j^n\|^2}{2\delta_p'^2}} \cdot e^{-\frac{\|r_i^n - r_j^n\|^2}{2\delta_r'^2}} \cdot e^{-\frac{\|y_i^n - y_j^n\|^2}{2\delta_y'^2}}, \tag{12}$$

where p_i and p_j denote the position, r_i and r_j denote the RGB information, y_i and y_j denote the predicted saliency values in the current frame, and $\{\delta_p', \delta_r', \delta_y'\}$ are the hyper-parameters for the scale of the three kinds of pair-wise similarities and are set to $\{6, 0.1, 0.1\}$. In Eq. (12), we also use shallow features to complement predicted saliency consistency.

The cross-frame saliency consistency function is defined as:

$$K_2'(i,j) = e^{-\frac{\|p_i^n - p_j^{n-1}\|^2}{2\delta_p'^2}} \cdot e^{-\frac{\|r_i^n - r_j^{n-1}\|^2}{2\delta_r'^2}} \cdot e^{-\frac{\|y_i^n - y_j^{n-1}\|^2}{2\delta_y'^2}}, \tag{13}$$

where r_j^{n-1} and y_j^{n-1} are the previous frame's RGB information and saliency predictions. With Eq. (13), the feature consistency can also be supervised with temporal information.

Overall, the cross-frame feature regularization and cross-frame saliency regularization heads in our CFMR can be reinforced by each other. Specifically, better saliency supervision provides a more accurate feature relationship for the

cross-frame saliency regularization head. On the other hand, a more accurate feature relationship facilitates to the production of better saliency predictions in turn. Thus, with such a CFMR training process, the network can learn more comprehensive object structure information for better final predictions.

3.4. Cross-frame Saliency Consistency Loss

To realize the cross-frame feature regularization in our cross-frame feature regularization head, we use the optical flow to map the previous frame to a pseudo current frame for temporal consistency. However, noise often exists in the optical flow, and the network will be misguided in the wrong direction. Therefore, we design a cross-frame saliency consistency (CSC) loss to regularize the training process based on the assumption that the predicted saliency results of the pseudo current frame and actual current frame should be consistent. Our CSC loss ($\mathcal{L}_{csc}$) can be defined as:

$$\mathcal{L}_{csc} = \frac{1}{M}\sum \gamma \frac{1 - \mathrm{SSIM}(y, \widetilde{y})}{2} + (1-\gamma)|y - \widetilde{y}|, \tag{14}$$

where y and $\widetilde{y}$ are the predicted saliency maps of the actual current frame and the pseudo current frame, respectively. M is the total number of pixels in one frame. SSIM$(\cdot)$ denotes the single scale SSIM [51] to measure the structure consistency of the two predicted saliency maps. γ is a hyper-parameter to balance our CSC loss. With Eq. (14), the network can encode more accurate temporal information for object structure information and generate better features of both the actual current frame and the pseudo current frame.

3.5. Loss Function

We follow the WSVSOD pipeline during the network training process [29, 30]. First, we pretrain the model on a static scribble annotated dataset S-DUTS [46], partial cross-entropy loss $\mathcal{L}_{pce}$ is used during this process, which can be mathematically expressed as follows:

$$\mathcal{L}_{pce} = \sum_{i \in \widetilde{\mathcal{Y}}} -\hat{y}_i \log y_i - (1-\hat{y}_i)\log(1-y_i), \tag{15}$$

where $\hat{y}$ denotes the groundtruth, y is the predicted values and $\widetilde{\mathcal{Y}}$ is the set of labeled pixels via scribble annotations. To learn appearance-sensitive features, two additional losses: gated structure-aware loss $\mathcal{L}_g$ and edge loss $\mathcal{L}_e$ proposed in [46] are also used. The pretrain loss $\mathcal{L}_{pretrain}$ is a combination of them:

$$\mathcal{L}_{pretrain} = \mathcal{L}_{pce} + \mathcal{L}_g + \mathcal{L}_e, \tag{16}$$

Then, the network will be finetuned on scribble-labeled video datasets, namely DAVIS-S and DAVSOD-S [29]. The overall loss $\mathcal{L}_{finetune}$ is defined as:

$$\mathcal{L}_{finetune} = \mathcal{L}_{pretrain} + \mathcal{L}_{cfmr}. \tag{17}$$

Note here, $\mathcal{L}_{cfmr}$ is the final loss function of our CFMR training process and is defined as:

$$\mathcal{L}_{cfmr} = \lambda_1 \mathcal{L}_{cfr} + \lambda_2 \mathcal{L}_{csr} + \lambda_3 \mathcal{L}_{csc}, \tag{18}$$

where λ_1, λ_2, λ_3 are loss weights.

4. Experiment

4.1. Implementation Details

We use the previous scribble-annotated WSVSOD prediction network in [29] as our backbone. The hyper-parameters $\{\lambda 1, \lambda 2, \lambda 3\}$ in Eq. (18) are set to $\{0.05, 0.01, 0.05\}$, respectively. The kernel sizes in our CFR loss and CSR loss are set to 5. The $\gamma = 0.85$ in Eq. (14) and the threshold to choose confidence pixels $\{\tau_f, \tau_g\}$ are set to $\{0.9, 0.1\}$. We use Adam as an optimizer to pretrain our model for 30 epochs first and then finetune our model for 25 epochs. The learning rate is set to 1e-4 for both the pretrain and finetune processes. The batch size is set to one, and the length of frames per batch is set to four. We conduct training on DAVIS-S [29] and DAVSOD-S [29]. Subsequently, to assess the generalization capability of our proposed method, we performed testing on six publicly available datasets separately, namely VOS [52], DAVIS [53], DAVSOD [12], FBMS [54], SegV2 [32] and ViSal [55]. During testing, we uniformly resize each frame to 256 $\times$ 256

Table 1: Comparison with previous state-of-the-art methods on six widely used benchmarks. The best results in weakly supervised methods are marked in red and the second-best ones in blue. ↑ & ↓ denote that larger and smaller are better, respectively. '*' denotes DenseCRF is utilized in the method. The epoch for the VOS dataset is set to 20 since it will overfit the noise caused by the low quality of optical flow.

		Fully Supervised Methods											Weakly Supervised Methods						
	Metric	EGNet	SCRN	PoolNet	FCNS	PDB	MGA	RCRNet	SSAV	PCSA	TENet	DCFNET	SSOD	GF	SAG	WSVSOD	CFMR	MPLA−Net*	CFMR*
		[56]	[57]	[58]	[39]	[36]	[10]	[59]	[12]	[60]	[11]	[61]	[46]	[55]	[62]	[29]	(Ours)	[41]	(Ours)
SegV2	$S_\alpha\uparrow$	0.845	0.817	0.782	-	0.864	0.880	0.843	0.849	0.866	0.868	0.893	0.733	0.699	0.719	0.804	0.831	0.836	0.848
	$F_\beta\uparrow$	0.774	0.760	0.704	-	0.808	0.829	0.782	0.797	0.811	0.810	0.837	0.664	0.592	0.634	0.738	0.786	0.778	0.802
	MAE↓	0.024	0.025	0.025	-	0.024	0.027	0.035	0.023	0.024	0.025	0.014	0.039	0.091	0.081	0.033	0.025	0.032	0.022
DAVIS	$S_\alpha\uparrow$	0.829	0.879	0.854	0.794	0.882	0.910	0.886	0.892	0.902	0.905	0.914	0.795	0.688	0.676	0.828	0.851	0.855	0.867
	$F_\beta\uparrow$	0.768	0.847	0.815	0.708	0.855	0.892	0.848	0.860	0.880	0.881	0.900	0.734	0.569	0.515	0.779	0.814	0.814	0.828
	MAE↓	0.057	0.029	0.038	0.061	0.028	0.023	0.027	0.028	0.022	0.017	0.016	0.044	0.100	0.103	0.037	0.031	0.033	0.028
DAVSOD	$S_\alpha\uparrow$	0.719	0.745	0.702	0.657	0.698	0.741	0.741	0.755	0.741	0.779	0.755	0.672	0.553	0.565	0.705	0.720	0.723	0.737
	$F_\beta\uparrow$	0.604	0.652	0.592	0.521	0.572	0.643	0.654	0.659	0.656	0.697	0.660	0.556	0.334	0.37	0.605	0.626	0.658	0.646
	MAE ↓	0.101	0.085	0.089	0.129	0.116	0.083	0.087	0.084	0.086	0.070	0.074	0.101	0.167	0.184	0.103	0.089	0.091	0.085
FBMS	$S_\alpha\uparrow$	0.878	0.876	0.839	0.794	0.851	0.908	0.872	0.879	0.868	0.916	-	0.747	0.651	0.659	0.778	0.783	-	0.798
	$F_\beta\uparrow$	0.848	0.861	0.830	0.759	0.821	0.903	0.859	0.865	0.837	0.915	-	0.727	0.571	0.564	0.786	0.787	-	0.802
	MAE↓	0.044	0.039	0.060	0.091	0.064	0.027	0.053	0.040	0.040	0.024	-	0.083	0.160	0.161	0.072	0.069	-	0.065
ViSal	$S_\alpha\uparrow$	0.946	0.948	0.902	0.881	0.907	0.940	0.922	0.942	0.946	0.949	0.952	0.853	0.757	0.749	0.857	0.864	0.890	0.883
	$F_\beta\uparrow$	0.941	0.946	0.891	0.852	0.888	0.936	0.907	0.938	0.941	0.949	0.953	0.831	0.683	0.688	0.831	0.848	0.881	0.869
	MAE↓	0.015	0.017	0.025	0.048	0.032	0.017	0.026	0.021	0.017	0.012	0.010	0.038	0.107	0.105	0.041	0.039	0.033	0.035
VOS	$S_\alpha\uparrow$	0.793	0.825	0.773	0.76	0.818	0.791	0.873	0.786	0.828	0.845	-	0.682	0.615	0.619	0.750	0.751	0.768	0.768
	$F_\beta\uparrow$	0.698	0.749	0.709	0.675	0.742	0.734	0.833	0.704	0.747	0.781	-	0.648	0.506	0.482	0.666	0.677	0.721	0.696
	MAE↓	0.082	0.067	0.082	0.099	0.078	0.075	0.051	0.091	0.065	0.052	-	0.106	0.162	0.172	0.091	0.092	0.084	0.088

and then feed it to the model to predict the final saliency maps without any post-processing following previous work [29]. Additionally, we also provide the results utilizing the DenseCRF [45] method in Tab. 1 for comparison purposes. The average processing time for analyzing a single frame within a sequence is 0.041s. All the experiments are implemented on one 24G NVIDIA TITAN RTX.

4.2. Dataset and Evaluation Metrics

Datasets. We follow the WSVSOD [29, 30] pipeline during the network training process using the same datasets. The model is first pretrained on static scribble datasets S-DUTS [46], and then finetuned on DAVIS-S [29] and DAVSOD-S [29]. We evaluate our model on six popular benchmarks: DAVSOD [12], VOS [52], DAVIS [53], ViSal [55], SegV2 [32], FBMS [54] to verify our method.

Evaluation Metrics. We use three metrics to evaluate the model: structure-based metric S_α, maximum F-measure (F_β), and Mean Absolute Error (MAE).

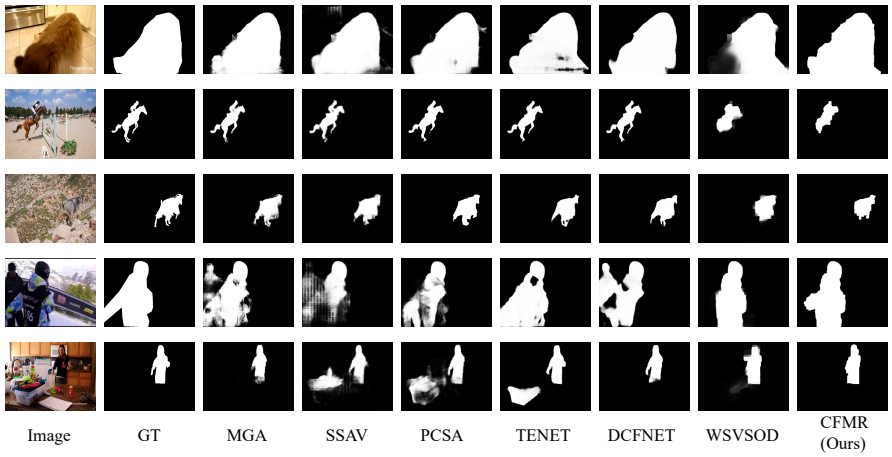

Figure 3: Qualitative comparison with different methods. MGA [10], SSAV [12], PCSA [60], TENet [11], DCFNET [61] are fully supervised methods, WSVSOD [29] and our CFMR are weakly supervised methods.

S_α [63] focuses on evaluating the structure of saliency maps and is defined as:

$$S_\alpha = \eta S_o + (1 - \eta) S_r, \tag{19}$$

where η is usually set to 0.5, S_o is object-aware similarity and S_r is region-aware similarity [63]. The F-measure considers both precision and recall, which are combined as:

$$F_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{20}$$

where β^2 is set to 0.3 [64].

Quantitative Comparison. In Tab. 1, we present the results of our method and the benchmark methods. It is evident that our method outperforms all other weakly supervised or unsupervised methods by a large margin on six datasets, except for MPLA-Net [41]. However, it is important to note that MPLA-Net [41] incorporates post-processing techniques, such as DenseCRF [45] and utilizes additional models such as the Itti model [42], CAM [43], super-pixel [44], to generate refined pseudo labels and eliminate noisy regions. In contrast, our method does not rely on any post-processing techniques or does

not employ additional models. Notably, when we integrate the DenseCRF [45] post-processing approach, our method achieves superior results compared to MPLA-Net on most datasets as illustrated in Tab. 1. Moreover, MPLA-Net is directly trained using five video datasets, namely SegV2 [32], ViSal [55], DAVIS [53], VOS [52], and DAVSOD [12], whereas our model undergoes initial pretraining on the image dataset S-DUTS [46] followed by finetuning on two video datasets, namely DAVIS-S [29] and DAVSOD-S [29] following previous work [29]. Therefore, the domain gap between these video datasets affects the results in ViSal and VOS of our method. Our method also demonstrates competitive performance even compared to some fully supervised methods, such as FCNS [39], EGNet [56] on DAVIS and DAVSOD datasets.

Qualitative Comparison. We demonstrate some visualization samples of our predicted video saliency maps in Fig. 3. We choose five representative frames, including various scenes from two popular datasets, DAVIS [53] and DAVSOD [12]. It is obvious that ours are smoother than the previous best weakly supervised method WSVSOD [29], even in the case of noisy background like row 3 in Fig. 3. Moreover, our method can predict better video saliency maps than previous state-of-the-art fully supervised methods (SSAV [12], PCSA [60], TENET [11]) in the case of complicated scenes like row 4 and 5 in Fig. 3.

4.3. Ablation Studies

In this section, we conduct our ablation studies on the DAVIS and DAVSOD datasets and use training with only $\mathcal{L}_{pretrain}$ as our baseline. In Tab. 2, we first conduct experiments on different loss functions of our method. Our method obtains the best performance using all the losses compared to the baseline. As listed in Tab. 2, each proposed loss function is crucial for our CFMR. Our $\mathcal{L}_{cfr}$ can obtain a gain of 1.3% for S_α and 2.4% for F_β compared to the baseline and $\mathcal{L}_{csr}$ can obtain a gain of 0.9% for S_α and 1.6% for F_β compared to the baseline on DAVIS dataset. Then, a combination of them can improve the performance by 2.1% for S_α and 3.5% for F_β. Finally, with the help of our $\mathcal{L}_{csc}$, we can obtain the performance of 0.851 for S_α and 0.814 for F_β, which is the new

Table 2: Ablation study for the influence of different loss functions. 'Base.' means using baseline.

Base.	$\mathcal{L}_{cfr}$	$\mathcal{L}_{csr}$	$\mathcal{L}_{csc}$	DAVIS		DAVSOD	
				S_α	F_β	S_α	F_β
✓				0.827	0.772	0.701	0.599
✓	✓			0.840	0.796	0.713	0.618
✓		✓		0.836	0.788	0.708	0.607
✓	✓	✓		0.848	0.807	0.717	0.622
✓	✓	✓	✓	**0.851**	**0.814**	**0.720**	**0.626**

Table 3: Analysis of performances with different consistency parts of $\mathcal{L}_{cfr}$ and $\mathcal{L}_{csr}$ losses on DAVIS and DAVSOD datasets.

Base.	Cur-fra	Cro-fra	DAVIS		DAVSOD	
			S_α	F_β	S_α	F_β
✓			0.827	0.772	0.701	0.599
✓	✓		0.837	0.791	0.710	0.613
✓		✓	0.834	0.785	0.706	0.609
✓	✓	✓	**0.840**	**0.796**	**0.713**	**0.618**

(a) Ablation study for different parts of $\mathcal{L}_{cfr}$ on DAVIS and DAVSOD datasets. 'Base.' means using baseline. 'Cur-fra' means using only the current-frame consistency in $\mathcal{L}_{cfr}$ and 'Cro-fra' means using only the cross-frame consistency in $\mathcal{L}_{cfr}$.

Base.	Cur-fra	Cro-fra	DAVIS		DAVSOD	
			S_α	F_β	S_α	F_β
✓			0.827	0.772	0.701	0.599
✓	✓		0.833	0.783	0.707	0.604
✓		✓	0.831	0.779	0.705	0.605
✓	✓	✓	**0.836**	**0.788**	**0.708**	**0.607**

(b) Ablation study for different parts of $\mathcal{L}_{csr}$ on DAVIS and DAVSOD datasets. 'Base.' means using baseline. 'Cur-fra' means using only the current-frame consistency in $\mathcal{L}_{csr}$ and 'Cro-fra' means using only the cross-frame consistency in $\mathcal{L}_{csr}$.

state-of-the-art performance and such performance illustrates the effectiveness of our CFMR. The results obtained on the DAVSOD dataset further illustrate the similar effectiveness of various loss functions employed in our method.

Impact of different consistency parts on CFR and CSR losses. In Tab. 3a, we conduct an ablation study for the influence of the current-frame and cross-frame parts of our CFR loss compared to the baseline. Specifically, the performance is increased by 1.0% for S_α and 1.9% for F_β with just the current-frame part on the DAVIS dataset. On the other hand, with only the cross-frame part, there are increments of about 0.7% for S_α and 1.3% for F_β on the DAVIS dataset. This phenomenon can prove the effectiveness of these

Table 4: Analysis of performances with different regularization information of $\mathcal{L}_{cfr}$ and $\mathcal{L}_{csr}$ losses on DAVIS and DAVSOD datasets.

Base.	Shallow	Deep	DAVIS		DAVSOD	
			S_α	F_β	S_α	F_β
✓			0.827	0.772	0.701	0.599
✓	✓		0.836	0.790	0.709	0.614
✓		✓	0.833	0.782	0.705	0.608
✓	✓	✓	**0.840**	**0.796**	**0.713**	**0.618**

(a) Ablation study for the impact of different regularization information on $\mathcal{L}_{cfr}$ loss. 'Base.' means using baseline. 'Shallow' means using shallow features. 'Deep' means using deep features.

Base.	Shallow	Saliency	DAVIS		DAVSOD	
			S_α	F_β	S_α	F_β
✓			0.827	0.772	0.701	0.599
✓	✓		0.833	0.785	0.706	0.604
✓		✓	0.830	0.781	0.704	0.604
✓	✓	✓	**0.836**	**0.788**	**0.708**	**0.607**

(b) Ablation study for the impact of different regularization information on $\mathcal{L}_{csr}$. 'Base.' means using baseline. 'Shallow' means using shallow feature values in $\mathcal{L}_{csr}$. 'Saliency' means using saliency predictions in $\mathcal{L}_{csr}$.

two parts. Finally, combining these two parts can improve the final results to 0.840 for S_α and 0.796 for F_β, which verifies that the two parts can leverage each other for better performance. The results achieved through the DAVSOD experiments exhibit a similar trend to those derived from the DAVIS dataset, which further reinforces the effectiveness of the components encompassing the $\mathcal{L}_{cfr}$ loss, namely the current-frame and cross-frame consistency parts.

We can acquire a similar trend of results in CSR loss from Tab. 3b, the performance is increased by 0.6% for S_α and 1.1% for F_β with just the current-frame part on the DAVIS dataset. On the other hand, with only the cross-frame part, there are increments of about 0.4% for S_α and 0.7% for F_β on the DAVIS dataset. This phenomenon can prove the effectiveness of these two parts in CSR loss. Finally, combining these two parts can improve the final results to 0.836 for S_α and 0.788 for F_β, which verifies that the two parts can leverage each other for better performance. The results from DAVSOD dataset provide further evidence supporting the efficacy of both the current-frame and cross-frame consistency components within the $\mathcal{L}_{csr}$ loss function.

Impact of different regularization information on $\mathcal{L}_{cfr}$ and $\mathcal{L}_{csr}$ losses. Tab. 4a shows the influence of shallow and deep features in our CFR loss. It can be found that shallow features can bring increments of 0.9% for S_α

Table 5: Ablation study for SSIM in the $\mathcal{L}_{csc}$ loss.

SSIM	DAVIS		DAVSOD	
	S_α	F_β	S_α	F_β
w/o	0.844	0.814	0.717	0.625
w/	**0.851**	0.814	**0.720**	**0.626**

and 1.8% for F_β on the DAVIS dataset. Deep features can improve by 0.6% for S_α and 1.0% for F_β on the DAVIS dataset. Finally, their combination acquires the highest results, illustrating that the balance of shallow and deep features can provide better regularization for WSVSOD. Moreover, the results obtained from the DAVSOD dataset offer additional support for the effectiveness of both shallow and deep features in the $\mathcal{L}_{cfr}$ loss function.

We also conduct experiments on the influence of shallow feature and saliency values of our CSR loss by comparing with the results of adding different regularization information separately to the baseline as shown in Tab. 4b. It can be found that with only saliency prediction for regularization, the improvement is limited for S_α but can improve the F_β by 0.9%. This may be caused by the noise in the initial predictions, which may lead the network trained in the wrong direction. However, with the balance between shallow information and saliency predictions, we can obtain the highest results of 0.836 for S_α and 0.788 for F_β on the DAVIS dataset. This illustrates that leveraging shallow information and saliency predictions can help regularize the feature consistency to help the network mine more comprehensive target features. In addition, the results derived from the analysis of the DAVSOD dataset provide additional evidence indicating the effectiveness of shallow features and saliency values within the $\mathcal{L}_{csr}$ loss function.

Impact of SSIM. We also evaluate the impact of SSIM in the CSC loss by just taking SSIM away from Eq. (14), and the results in DAVIS and DAVSOD datasets are shown in Tab. 5. It can be seen that SSIM can help provide better prediction structural consistency.

Impact of kernel size. We conduct the ablation study on the choice of k in

Table 6: Ablation study for the impact of kernel size k in $\mathcal{L}_{cfr}$ loss and $\mathcal{L}_{csr}$ loss.

kernel size	S_α	F_β
3	0.838	0.801
5	**0.848**	**0.807**
7	0.842	0.804

Table 7: Analysis of performances with different thresholds in the $\mathcal{L}_{csr}$ loss on DAVIS.

τ_f	S_α	F_β
0.8	0.832	0.782
0.9	**0.836**	**0.788**
0.95	0.833	0.784

(a) Ablation study for τ_f in the $\mathcal{L}_{csr}$ loss.

τ_g	S_α	F_β
0.05	0.833	0.785
0.1	**0.836**	**0.788**
0.2	0.835	0.787

(b) Ablation study for τ_g in the $\mathcal{L}_{csr}$ loss.

our CFR loss and CSR loss in Tab. 6. We find that both a smaller k and a larger k will lead to a slight decline in performance compared with $k = 5$. Additionally, a larger k will take more computing resources. Thus, we choose $k = 5$ for our CFMR to aggregate sufficient consistency relationship with efficiency.

Impact of τ_f and τ_g. In the $\mathcal{L}_{csr}$ loss, τ_f and τ_g serve as threshold values that establish upper and lower bounds for selecting pixels with confident saliency values both in the foreground and background to help provide the cross-frame saliency consistency. The results are shown in Tab. 7a and Tab. 7b. In Tab. 7a, the highest result is obtained when τ_f is set to 0.9. A higher τ_f will filter out some pixels belonging to the saliency object and a lower τ_f may incorporate background noise. Both of them result in a decline in the performance. The results presented in Tab. 7b indicate the impact of τ_g in the $\mathcal{L}_{csr}$ loss. Compared to τ_f, τ_g is less sensitive, and the optimal performance is attained when τ_g is set to 0.1.

Limitation and Failure Cases. In this section, we discuss the limitations and failure cases of our proposed method as illustrated in Fig. 4. As can be seen from the figure, our method is capable of accurately locating the salient object in a video frame. However, in situations where there are occlusions,

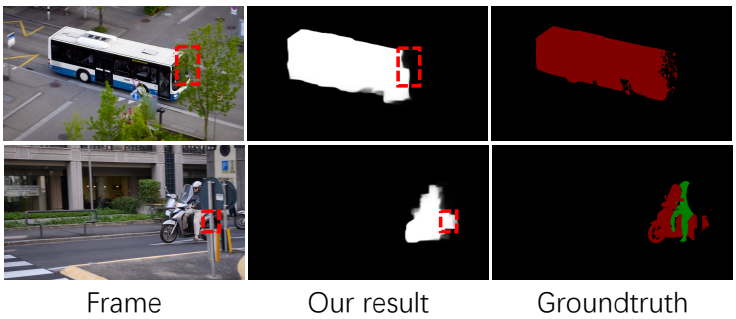

Figure 4: Visualization of failure cases. In occlusion situations, our method may encounter difficulties in precisely delineating the boundaries of occluded regions, which can result in inaccurate salient object prediction.

our method may encounter challenges in effectively delineating the boundary of occlusion positions. This is indeed a significant challenge in the VSOD task, and as a future direction, we intend to explore more approaches to alleviate such a problem.

5. Conclusion

In this paper, we propose a CFMR training process realized by a cross-frame feature regularization head and a cross-frame saliency regularization head to mitigate the issue of incomplete object structure caused by scribble annotations. A CFR loss is designed in our cross-frame feature regularization head to assist scribble annotations for better saliency predictions. Meanwhile, considering deep features fail to provide sufficient consistency due to scribble annotations, we design a CSR loss in our cross-frame saliency regularization head to promote feature consistency quality by using saliency maps as supervision. In this way, the mutual interaction of these two heads can reinforce the network to learn comprehensive object structure information in a more accurate direction. Extensive experiments illustrate our method outperforms the existing state-of-the-art WSVSOD method.

6. Acknowledge

This work was supported by the National Key R&D Program of China (No.2022YFE0200300), the National Natural Science Foundation of China (No. 61972323, 62331003), Suzhou Basic Research Program (SYG202316) and XJTLU REF-22-01-010, XJTLU AI University Research Centre, Jiangsu Province Engineering Research Centre of Data Science and Cognitive Computation at XJTLU and SIP AI innovation platform (YZCXPT2022103), Suzhou Municipal Key Laboratory for Intelligent Virtual Engineering (SZS2022004).

This work was also supported by the Taishan Scholar Program of Shandong (No. tsqn202306130), the Shandong Natural Science Foundation (Grant No. ZR2023QF046), Independent Innovation Research Project of China University of Petroleum (East China) (No.22CX06060A), Qingdao Postdoctoral Applied Research Project (No. DBSH20230102091).

References

[1] Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, Y. Yu, Multi-source weak supervision for saliency detection, in: CVPR, 2019, pp. 6074–6083.

[2] G. Li, Y. Xie, L. Lin, Weakly supervised salient object detection using image labels, in: AAAI, 2018, pp. 6074–6083.

[3] N. Liu, W. Zhao, D. Zhang, J. Han, L. Shao, Light field saliency detection with dual local graph learning and reciprocative guidance, in: ICCV, 2021, pp. 4712–4721.

[4] G. Wang, C. Chen, D. Fan, A. Hao, H. Qin, From semantic categories to fixations: A novel weakly-supervised visual-auditory saliency detection approach, in: CVPR, 2021, pp. 15119–15128.

[5] D. Zhang, H. Tian, J. Han, Few-cost salient object detection with adversarial-paced learning, in: NeurIPS, 2020, pp. 12236–12247.

[6] J. Zhang, J. Xie, N. Barnes, Learning noise-aware encoder-decoder from noisy labels by alternating back-propagation for saliency detection, in: ECCV, 2020, pp. 349–366.

[7] M. Zhang, T. Liu, Y. Piao, S. Yao, H. Lu, Auto-msfnet: Search multi-scale fusion network for salient object detection, in: ACM MM, 2021, pp. 667–676.

[8] Z. Wu, L. Su, Q. Huang, Cascaded partial decoder for fast and accurate salient object detection, in: CVPR, 2019, pp. 3907–3916.

[9] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, E. Ding, A mutual learning method for salient object detection with intertwined multi-supervision, in: CVPR, 2019, pp. 8150–8159.

[10] H. Li, G. Chen, G. Li, Y. Yu, Motion guided attention for video salient object detection, in: ICCV, 2019, pp. 7274–7283.

[11] S. Ren, C. Han, X. Yang, G. Han, S. He, Tenet: Triple excitation network for video salient object detection, in: ECCV, 2020, pp. 212–228.

[12] D.-P. Fan, W. Wang, M.-M. Cheng, J. Shen, Shifting more attention to video salient object detection, in: CVPR, 2019, pp. 8554–8564.

[13] C. Chen, S. Li, Y. Wang, H. Qin, A. Hao, Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion, IEEE TIP 26 (7) (2017) 3156–3170.

[14] Z. Liu, J. Li, L. Ye, G. Sun, L. Shen, Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation, IEEE TCSVT 27 (12) (2016) 2527–2542.

[15] H. Hadizadeh, I. V. Bajić, Saliency-aware video compression, IEEE TIP 23 (1) (2013) 19–33.

[16] L. Itti, Automatic foveation for video compression using a neurobiological model of visual attention, IEEE TIP 13 (10) (2004) 1304–1318.

[17] G.-P. Ji, K. Fu, Z. Wu, D.-P. Fan, J. Shen, L. Shao, Full-duplex strategy for video object segmentation, in: ICCV, 2021, pp. 4922–4933.

[18] Y. Lee, H. Seong, E. Kim, Iteratively selecting an easy reference frame makes unsupervised video object segmentation easier, in: AAAI, 2022, pp. 1245–1253.

[19] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, H. Xia, End-to-end video instance segmentation with transformers, in: CVPR, 2021, pp. 8741–8750.

[20] M. Sun, J. Xiao, E. G. Lim, Y. Xie, J. Feng, Adaptive roi generation for video object segmentation using reinforcement learning, Pattern Recognition 106 (2020) 107465.

[21] Y. Pan, T. Yao, H. Li, T. Mei, Video captioning with transferred semantic attributes, in: CVPR, 2017, pp. 6504–6512.

[22] L. Gao, Z. Guo, H. Zhang, X. Xu, H. T. Shen, Video captioning with attention-based lstm and semantic consistency, IEEE TMM 19 (9) (2017) 2045–2055.

[23] B. Zhang, J. Xiao, Y. Wei, K. Huang, S. Luo, Y. Zhao, End-to-end weakly supervised semantic segmentation with reliable region mining, Pattern Recognition 128 (2022) 108663.

[24] B. Zhang, J. Xiao, Y. Wei, M. Sun, K. Huang, Reliability does matter: An end-to-end weakly supervised semantic segmentation approach, in: AAAI, 2020, pp. 12765–12772.

[25] B. Zhang, J. Xiao, J. Jiao, Y. Wei, Y. Zhao, Affinity attention graph neural network for weakly supervised semantic segmentation, IEEE TPAMI 44 (2021) 8082–8096.

[26] L. Yu, W. Xiang, J. Fang, Y.-P. P. Chen, L. Chi, ex-vit: A novel explainable vision transformer for weakly supervised semantic segmentation, Pattern Recognition 142 (2023) 109666.

[27] H. Qin, W. Xie, Y. Li, K. Jiang, J. Lei, Q. Du, Weakly supervised adversarial learning via latent space for hyperspectral target detection, Pattern Recognition 135 (2023) 109125.

[28] X. Liu, Q. Yuan, Y. Gao, K. He, S. Wang, X. Tang, J. Tang, D. Shen, Weakly supervised segmentation of covid19 infection with scribble annotation on ct images, Pattern recognition 122 (2022) 108341.

[29] W. Zhao, J. Zhang, L. Li, N. Barnes, N. Liu, J. Han, Weakly supervised video salient object detection, in: CVPR, 2021, pp. 16826–16835.

[30] S. Gao, H. Xing, W. Zhang, Y. Wang, Q. Guo, W. Zhang, Weakly supervised video salient object detection via point supervision, in: ACM MM, 2022, pp. 3656–3665.

[31] S. Yu, B. Zhang, J. Xiao, E. G. Lim, Structure-consistent weakly supervised salient object detection with local saliency coherence, in: AAAI, 2021, pp. 3234–3242.

[32] F. Li, T. Kim, A. Humayun, D. Tsai, J. M. Rehg, Video segmentation by tracking many figure-ground segments, in: ICCV, 2013, pp. 2192–2199.

[33] C. Chen, G. Wang, C. Peng, Y. Fang, D. Zhang, H. Qin, Exploring rich and efficient spatial temporal interactions for real-time video salient object detection, IEEE TIP 30 (1) (2021) 3995–4007.

[34] C. Chen, H. Wang, Y. Fang, C. Peng, A novel long-term iterative mining scheme for video salient object detection, IEEE TCSVT 32 (2022) 7662–7676.

[35] G. Li, Y. Xie, T. Wei, K. Wang, L. Lin, Flow guided recurrent neural encoder for video salient object detection, in: CVPR, 2018, pp. 3243–3252.

[36] H. Song, W. Wang, S. Zhao, J. Shen, K.-M. Lam, Pyramid dilated deeper convlstm for video salient object detection, in: ECCV, 2018, pp. 715–731.

[37] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, A. Borji, Revisiting video saliency prediction in the deep learning era, IEEE TPAMI 43 (1) (2019) 220–237.

[38] P. Chen, J. Lai, G. Wang, H. Zhou, Confidence-guided adaptive gate and dual differential enhancement for video salient object detection, in: ICME, 2021, pp. 1–6.

[39] W. Wang, J. Shen, L. Shao, Video salient object detection via fully convolutional networks, IEEE TIP 27 (1) (2017) 38–49.

[40] N. Liu, K. Nan, W. Zhao, X. Yao, J. Han, Learning complementary spatial–temporal transformer for video salient object detection, IEEE TNNLS (2023) 1–11.

[41] C. Ma, L. Du, L. Zhuo, J. Li, Mpla-net: Multiple pseudo label aggregation network for weakly supervised video salient object detection, IEEE TCSVT (Early Access).

[42] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE TPAMI 20 (11) (1998) 1254–1259.

[43] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: CVPR, 2016, pp. 2921–2929.

[44] Y. Piao, J. Wang, M. Zhang, H. Lu, Mfnet: Multi-filter directive network for weakly supervised salient object detection, in: ICCV, 2021, pp. 4136–4145.

[45] P. Krähenbühl, V. Koltun, Efficient inference in fully connected crfs with gaussian edge potentials, in: NeurIPS, 2011, pp. 834–848.

[46] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, Y. Dai, Weakly-supervised salient object detection via scribble annotations, in: CVPR, 2020, pp. 12546–12555.

[47] Z. Huang, T.-Z. Xiang, H.-X. Chen, H. Dai, Scribble-based boundary-aware network for weakly supervised salient object detection in remote sensing images, ISPRS Journal of Photogrammetry and Remote Sensing 191 (2022) 290–301.

[48] S. Gao, W. Zhang, Y. Wang, Q. Guo, C. Zhang, Y. He, W. Zhang, Weakly-supervised salient object detection using point supervison, in: AAAI, 2022, pp. 670–678.

[49] A. Li, Y. Mao, J. Zhang, Y. Dai, Mutual information regularization for weakly-supervised rgb-d salient object detection, IEEE TCSVT (Early Access).

[50] Z. Teed, J. Deng, Raft: Recurrent all-pairs field transforms for optical flow, in: ECCV, 2020, pp. 402–419.

[51] C. Godard, O. Mac Aodha, G. J. Brostow, Unsupervised monocular depth estimation with left-right consistency, in: CVPR, 2017, pp. 270–279.

[52] J. Li, C. Xia, X. Chen, A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection, IEEE TIP 27 (1) (2017) 349–364.

[53] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, in: CVPR, 2016, pp. 724–732.

[54] P. Ochs, J. Malik, T. Brox, Segmentation of moving objects by long term video analysis, IEEE TPAMI 36 (6) (2013) 1187–1200.

[55] W. Wang, J. Shen, L. Shao, Consistent video saliency using local gradient flow optimization and global refinement, IEEE TIP 24 (11) (2015) 4185–4196.

[56] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, M.-M. Cheng, Egnet: Edge guidance network for salient object detection, in: ICCV, 2019, pp. 8779–8788.

[57] Z. Wu, L. Su, Q. Huang, Stacked cross refinement network for edge-aware salient object detection, in: ICCV, 2019, pp. 7264–7273.

[58] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, J. Jiang, A simple pooling-based design for real-time salient object detection, in: CVPR, 2019, pp. 3917–3926.

[59] P. Yan, G. Li, Y. Xie, Z. Li, C. Wang, T. Chen, L. Lin, Semi-supervised video salient object detection using pseudo-labels, in: ICCV, 2019, pp. 7284–7293.

[60] Y. Gu, L. Wang, Z. Wang, Y. Liu, M.-M. Cheng, S.-P. Lu, Pyramid constrained self-attention network for fast video salient object detection, in: AAAI, 2020, pp. 10869–10876.

[61] M. Zhang, J. Liu, Y. Wang, Y. Piao, S. Yao, W. Ji, J. Li, H. Lu, Z. Luo, Dynamic context-sensitive filtering network for video salient object detection, in: ICCV, 2021, pp. 1553–1563.

[62] W. Wang, J. Shen, F. Porikli, Saliency-aware geodesic video object segmentation, in: CVPR, 2015, pp. 3395–3402.

[63] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, A. Borji, Structure-measure: A new way to evaluate foreground maps, in: ICCV, 2017, pp. 4548–4557.

[64] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, Q. Huang, Review of visual saliency detection with comprehensive information, IEEE TCSVT 29 (10) (2018) 2941–2959.