

SCRIBBLE-SUPERVISED RGB-T SALIENT OBJECT DETECTION

Anonymous ICME submission

ABSTRACT

Salient object detection segments attractive objects in scenes. RGB and thermal modalities provide complementary information and scribble annotations alleviate large amounts of human labor. Based on the above facts, we propose a scribble-supervised RGB-T salient object detection model. By a four-step solution (expansion, prediction, aggregation, and supervision), label-sparse challenge of scribble-supervised method is solved. To expand scribble annotations, we collect the superpixels that foreground scribbles pass through in RGB and thermal images, respectively. The expanded multi-modal labels provide the coarse object boundary. To further polish the expanded labels, we propose a prediction module to alleviate the sharpness of boundary. To play the complementary roles of two modalities, we combine the two into aggregated pseudo labels. Supervised by scribble annotations and pseudo labels, our model achieves the state-of-the-art performance on the relabeled RGBT-S dataset. Furthermore, the model is applied to RGB-D and video scribble-supervised applications, achieving consistently excellent performance.

Index Terms— multi-modal, scribble annotation, salient object detection, superpixel, weakly supervised learning

1. INTRODUCTION

Background. In the cluttered background, low-illumination, and foggy weather scenes, the thermal image is a promising supplement to the RGB image because it is insensitive to lighting and weather by capturing the radiated heat of objects. Therefore, RGB-T salient object detection (SOD) which segments the attractive objects shows promising research prospect.

Challenge. Scribble annotations can alleviate the labeling cost and have been widely used in remote sensing image [1, 2], medical image [3, 4], and 3D image segmentation [5]. However, object structures and boundary details cannot be easily inferred [6]. Especially, it is more difficult to extract the complete objects from low-quality RGB images.

Existing Solutions. Some methods [7, 8, 9] use edge detection to provide more structure information. Some methods [1, 4] use simple linear iterative clustering (SLIC) [10] to collect the superpixels in RGB images that scribbles pass through, generating pseudo labels for supervision. Some methods [7, 9] solve the challenge of low-quality RGB images by appending depth and optical flow information.

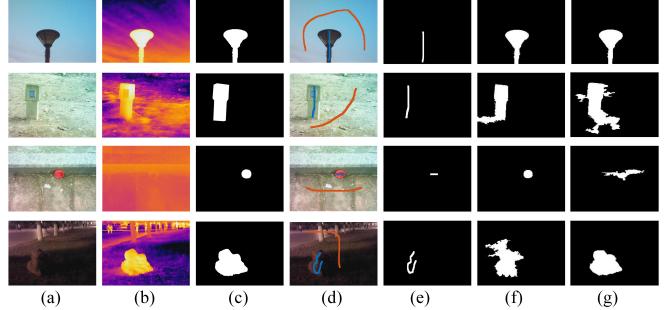


Fig. 1: Scribble-supervised RGB-T SOD datasets and corresponding multi-modal expanded labels. (a) RGB images (b) thermal images (c) ground truth (d) scribble annotations (red: background; blue: foreground) (e) foreground scribbles (f) RGB expanded labels (g) thermal expanded labels.

Our Findings. Scribble annotations describe the partial cues of objects without complete boundary, while superpixel segmentation can provide the coarse object boundary according to the local consistency in features. As a result, according to superpixels, we expand the foreground scribble annotations to generate the expanded labels from the two modalities (RGB and Thermal). By the observations from Fig 1 (f-g), we find that expanded labels provide abundant and useful information(1st row). Furthermore, we also find that expanded labels are inaccurate due to some noisy boundaries (2nd row). Most importantly, the complementarity of two modalities is disclosed that the expanded labels from one modality are inaccurate while those from the other modality match the salient objects (3rd and 4th rows). Accordingly, besides original scribble supervision, we append the supervision of the expanded labels. However, the expanded labels are not always accurate due to the limitations of the superpixel segmentation, which will affect the final prediction.

Our Solution. The four-step solution, *Expansion, Prediction, Aggregation, and Supervision*, is proposed to solve the above problem. Specifically, we first use RGB and thermal superpixels to expand foreground scribble annotations, generating the multi-modal expanded labels. Then, we propose a prediction module to predict the expanded labels from two modalities, respectively. Furthermore, the predicted multi-modal labels are averaged and smoothed to form the aggregated pseudo labels which can be further used to supervise the model training.

Our Contributions.

- A scribble-supervised RGB-T SOD model is proposed

- based on our launched scribble-supervised RGB-T-S dataset, achieving the comparable performance with the state-of-the-art fully-supervised methods. Multi-modal superpixel expansion obtains the completeness of objects by the combination of RGB and thermal images, so that the excellent pseudo labels are generated.
- Different from existing superpixel expansion models (ScRoadExtractor [1], Scribble2D5 [4]), we propose to supervise the model via the predicted pseudo labels instead of direct expanded ones.
 - Different from existing multi-modal scribble-supervised SOD models (RGB-D DENet [7], Video WSVOD [8]), we achieve multi-modal complementarity exploration in the supervision, rather than only in the model structures.

2. PROPOSED METHOD

2.1. Overview

The model (Fig 2) builds on a baseline network supervised by scribble annotations. Then, two foreground scribble expansion modules are used to expand the annotations via the superpixels, generating the expanded multi-modal labels. Then, the expanded labels supervise the prediction modules to generate the predicted labels. Next, the predicted labels are aggregated into the pseudo labels which will additionally supervise the baseline model.

2.2. Scribble-Supervised Baseline Model

In the baseline network shown in the green background of Fig 2, RGB image X and thermal image T are fed into two encoders to extract multi-modal features $F^x = \{f_i^x\}_{i=1}^4$ and $F^t = \{f_i^t\}_{i=1}^4$, respectively, where i denotes the number of layers. Then these features are concatenated and fed into a conventional saliency decoder which consists of adjacent layer concatenation, convolution, and upsampling operations to generate the result saliency map $R = \{r_i\}_{i=1}^N$ which is further supervised by the scribble annotation $Y = \{y_i\}_{i=1}^N$, where $y_i = 1/2/0$ indicates foreground, background, and unknown pixels, respectively, and N is the number of pixels in each image.

$$\begin{aligned} F^x &= \text{Encoder}(X) \\ F^t &= \text{Encoder}(T) \\ R &= \text{Decoder}(\text{Concat}(F^x, F^t)) \end{aligned} \quad (1)$$

where $\text{Concat}(\cdot)$ denotes concatenation operation.

Inspired by WSSA [6] and SCWSSOD [11], we use (1) partial cross entropy loss (L_{pce}) to supervise predicted saliency pixels only corresponding to scribble annotations; (2) local saliency coherence loss (L_{lsc}) to enforce similar pixels in the adjacent region to share consistent saliency scores; (3) smoothness loss (L_{sl}) to guarantee the local smoothness and

salient distinction along image edges; (4) structure consistency loss (L_{ssc}) to enforce the consistency on saliency maps from different input scales.

The scribble supervision loss $L_{scribble}$ is given by the combination of the above loss functions:

$$L_{scribble} = L_{pce} + L_{lsc} + L_{sl} + L_{ssc} \quad (2)$$

The loss uses the sparse annotations which indicate the foreground and background and adds the local and scale constraint. However, all the cues come from RGB images. When the quality of RGB images are low, the performance will be bad. Next, we will elaborate on the additional pseudo label supervision from RGB and thermal modalities.

2.3. Pseudo Label Generation and Supervision

Existing methods [1, 4] generate expanded labels by collecting the superpixels that scribbles pass through in RGB images. In the RGB-T SOD task, the expanded labels can be obtained from RGB images and thermal images. From the observation of Fig 1, we find the inaccuracy and complementarity of multi-modal expanded labels. Accordingly, we attempt to predict the pseudo labels from the two modalities via the supervision of expanded labels and then aggregate them to form the final pseudo labels, rather than directly use the expanded labels.

Our four-step solution includes expansion, prediction, aggregation, and supervision.

Expansion. A foreground scribble expansion module is proposed to generate expanded labels of two modalities. Specifically, the foreground scribbles are extracted from scribble annotations and the image is partitioned into multiple superpixels via SLIC. If the intersection of a superpixel and foreground scribble is not null, the superpixel is marked as the foreground. As a result, the scribble annotation Y is expanded via the RGB superpixels and thermal superpixels, generating corresponding expanded RGB labels $\hat{Y}^x = \{\hat{y}_i^x\}_{i=1}^N$ and expanded thermal labels $\hat{Y}^t = \{\hat{y}_i^t\}_{i=1}^N$. The algorithm of foreground scribble expansion module is shown in Alg 1.

Prediction. The expanded labels are obtained by SLIC which considers the local consistency. However, there are some noises in the boundary of objects, which can be shown in Fig 3 (c-d). Therefore, a prediction module is designed to predict the pseudo labels from each modality under the supervision of the expanded multi-modal labels.

$$\begin{aligned} P^x &= \text{Prediction}(f_4^x) \\ P^t &= \text{Prediction}(f_4^t) \end{aligned} \quad (3)$$

where $P^x = \{p_i^x\}_{i=1}^N$ and $P^t = \{p_i^t\}_{i=1}^N$ are predicted RGB and thermal labels.

The prediction module consists of successive convolution and upsampling operations conducted on the highest layer features of encoders, resulting more smoothed boundaries, as

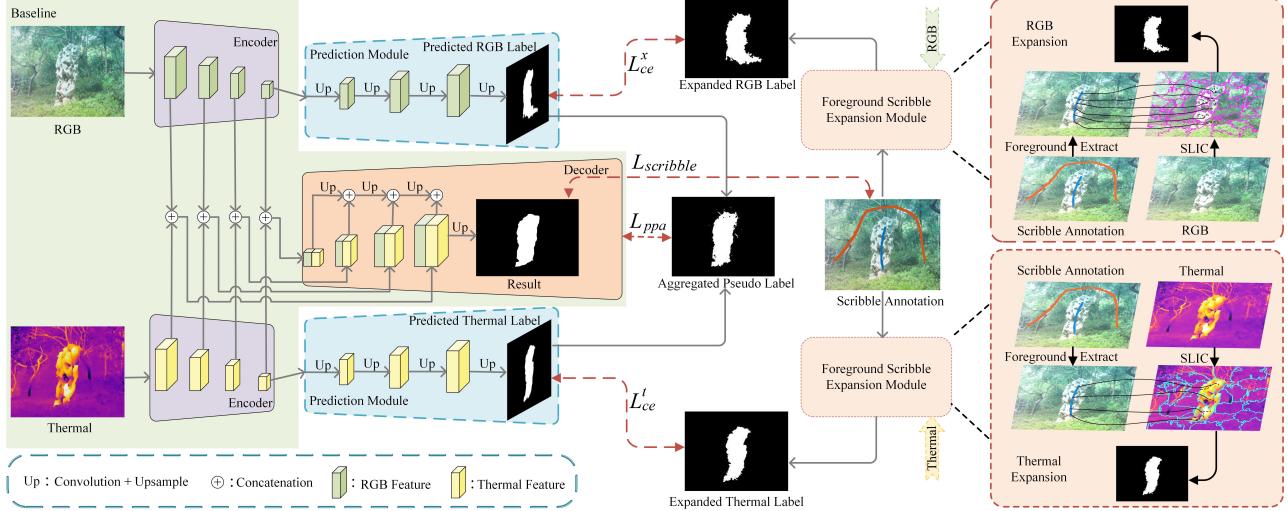


Fig. 2: The framework of the proposed model.

Algorithm 1: Foreground Scribble Expansion.

Input: (1) An RGB or thermal image M ;
(2) Scribble annotation $Y = \{y_i\}_{i=1}^N$, where $y_i = 1/2/0$ indicates foreground, background, and unknown pixels, respectively, and N is the number of pixels;

Output: Expanded labels \hat{Y} ;

/*Step1: Use SLIC algorithm to segment the image M , generating the superpixel image mask set $\mathbb{P} = \{P^k\}_{k=1}^T$, where T is the number of superpixels; Each P^k only highlights the current superpixel.*/

- 1: $\mathbb{P} \leftarrow SLIC(M)$;
- /*Step2: Extract the foreground scribble $FS = \{fs_i\}_{i=1}^N$ from scribble annotation Y .*/
- 2: Initialize $FS \leftarrow zeros(N)$;
- 3: **for** $i=1$ to N **do**
- 4: **if** $y_i = 1$ **then**
- 5: $fs_i \leftarrow 255$
- 6: **else**
- 7: $fs_i \leftarrow 0$
- 8: **end if**
- 9: **end for**
- /*Step3: Mark the superpixel as the foreground if the intersection of a superpixel and foreground scribble is not null.*/
- 10: Initialize $\hat{Y} \leftarrow zeros(N)$;
- 11: **for** each P^k in \mathbb{P} **do**
- 12: **if** $P^k \cap FS \neq \emptyset$ **then**
- 13: $\hat{Y} \leftarrow \hat{Y} + P^k$
- 14: **end if**
- 15: **end for**
- 16: **return** \hat{Y} .

shown in Fig 3 (e-f)vs.(c-d). The predicted labels are supervised by the expanded labels via cross entropy loss.

$$L_{ce}^x = -\frac{1}{N} \sum_i p_i^x \log \hat{y}_i^x + (1 - p_i^x) \log(1 - \hat{y}_i^x) \quad (4)$$

$$L_{ce}^t = -\frac{1}{N} \sum_i p_i^t \log \hat{y}_i^t + (1 - p_i^t) \log(1 - \hat{y}_i^t)$$

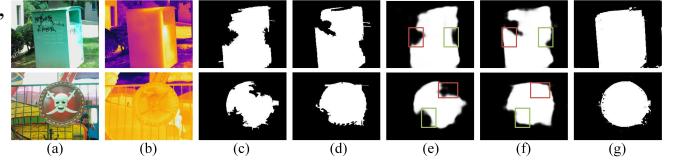


Fig. 3: Examples to show the advantage of predicted multi-modal labels compared with expanded multi-modal labels and the complementarity of multi-modal labels. (a-b) RGB and thermal images; (c-d) expanded RGB and thermal labels; (e-f) predicted RGB and thermal labels; (g) aggregated pseudo labels.

Aggregation. Two predicted labels are averaged and smoothed by original RGB images via PAR method [12] to generate aggregated pseudo labels $\bar{Y} = \{\bar{y}_i\}_{i=1}^N$. PAR incorporates the RGB and spatial information to define an affinity kernel measuring its proximity to its neighbours and iteratively applies this kernel to the average value via an adaptive convolution [13].

$$\bar{Y} = PAR(Avg(P^x, P^t), X) \quad (5)$$

Here, we only use RGB images instead of the thermal images to refine the average values because RGB images provide the more abundant information benefiting the refinement compared with thermal images.

From Fig 3 (g)vs.(e-f), we can observe that the aggregated pseudo labels are better via the mutual supplementation of two modalities.

Supervision. The supervision of aggregated pseudo labels are appended to the baseline by pixel position aware loss L_{ppa} .

$$L_{ppa} = L_{wbce}(R, \bar{Y}) + L_{wiou}(R, \bar{Y}) \quad (6)$$

where L_{wbce} and L_{wiou} are the weighted binary cross entropy loss and weighted IoU loss, respectively [14].

The final loss function L for training the proposed model is given by:

$$L = L_{scribble} + L_{ce}^x + L_{ce}^t + L_{ppa} \quad (7)$$

During testing, we only retain the baseline network and discard the prediction modules and foreground scribble expansion modules for acceleration.

3. EXPERIMENTS

3.1. Dataset and Evaluation Metrics

We relabel RGB-T SOD training set via scribbles, which includes 2,500 image pairs in VT5000 [15], namely RGBT-S dataset. The remaining samples of VT5000, the whole VT821 [16] and VT1000 [17] are used to be tested. VT821[16] contains 821 manually registered image pairs, so that the quality of some thermal infrared images are bad, for example, vacant regions or grey-scale map. VT1000[17] contains 1,000 RGB-T image pairs captured with highly aligned RGB and thermal cameras in relatively simple scenes. VT5000[15] contains 5,000 pairs of high-resolution, high-diversity, and low-deviation RGB-T images, involving all-day scenes. The evaluation metrics follow the BBSNet [18] and SwinNet [19].

3.2. Implementation Details

Our model is implemented based on PyTorch on a PC with a NVIDIA RTX 3090 GPU. The input image size is 256×256 . The number of superpixels is 40. Flipping, cropping, and rotating are used to augment the training set. We employ the Adam optimizer [20] to train our model. The max training epoch is set to 300 and batch size is set to 16. The initial learning rate is $5e-5$ and the learning rate will be divided by 10 every 200 epochs. The train process takes about 20 hours.

3.3. Comparison Experiments

Comparison Methods and Instructions. The fully-supervised multi-modal methods (BBSNet [18], MIDD [21], SwinNet [19], TNet [22]), scribble-supervised RGB methods (WSSA [6], SCWSSOD [11], SCOD [23]), scribble-supervised RGB-D method (DENet [7]), and the video ones (WSVOD [8]) are compared. For fairness, we use the source codes provided by the authors to re-train the RGBT-S datasets. For RGB based methods, multi-modal image pairs are first

concatenated and then fed into the models. Our method provides ResNet-50[24] and PVTv2-B2[25] version. All the evaluation metrics are calculated using the same evaluation code on result saliency maps. The win/tie/losses of fully-supervised methods show our PVT version vs. the other fully-supervised methods, while those of scribble-supervised methods mean our ResNet version vs. the other scribble-supervised ones.

Quantitative Analysis. From Table 1, we can find the below facts. Our ResNet model achieves the best results among the scribble-supervised methods. It fairly verifies the advantage of our model. Furthermore, our PVT model significantly outperforms the other scribble-supervised methods. Moreover, it is also comparable with fully-supervised methods which can be seen from their win/tie/loss values. It scientifically exhibits the superiority of the proposed model. The experimental data discloses the effectiveness of superpixel expansion of scribble annotations because the superpixel can provide a coarse boundary based on the given scribble. Especially, two-modal superpixel expansion can complement each other and play an approximate role as ground truth. Certainly, the proposed prediction module also polishes the coarse boundary of SLIC segmentation. It also improves the performance.

Visual Analysis. Fig 4 shows some visual examples in small objects, complex background, blur RGB images, multiple objects, connected objects, and cluttered scenes. Our results are obviously better, which indicate the advantage of our model. Certainly, there are some failure cases, for example, transparent, hollow, and thin objects shown in Fig 5. It can be inferred from the performances of the other scribble-supervised methods that these scenes will be the great challenge of future research.

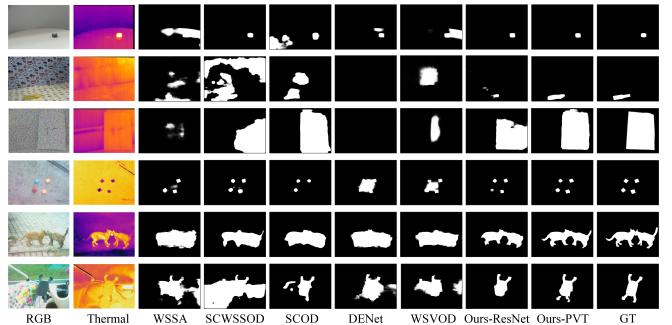


Fig. 4: Visual performance of scribble-supervised compared methods.

3.4. Ablation Study

Expanded vs. Predicted Ablation study about expanded and predicted labels are shown in Table 2. The performance under the appended supervision of aggregated pseudo labels which are averaged and smoothed by expanded multi-modal labels is better than the baseline. It shows the effectiveness

Table 1: Comparison with fully-supervised methods and scribble-supervised methods. The best results in each group are in bold.

Datasets	Metric	Fully Sup. Methods				Scribble Sup. Methods						
		BBSNet [18] ECCV20	MIDD [21] TIP21	SwinNet [19] TCSV22	TNet [22] TMM22	WSSA [6] CVPR20	SCWSSOD [11] AAAI21	SCOD [23] AAAI23	DENet [7] TIP22	WSVOD [8] CVPR21	Ours ResNet	Ours PVT
VT821 [16]	$S \uparrow$.862	.871	.904	.899	.805	.845	.822	.813	.822	.857	.895
	$F_\beta \uparrow$.762	.804	.847	.842	.722	.802	.772	.705	.755	.835	.875
	$E_\xi \uparrow$.879	.895	.926	.919	.848	.884	.879	.827	.868	.904	.934
	$M \downarrow$.046	.045	.030	.030	.058	.050	.048	.054	.052	.036	.027
VT1000 [17]	$S \uparrow$.926	.915	.938	.929	.882	.907	.880	.906	.886	.906	.925
	$F_\beta \uparrow$.847	.882	.896	.889	.842	.895	.849	.870	.863	.899	.914
	$E_\xi \uparrow$.920	.933	.947	.937	.907	.937	.920	.925	.916	.939	.949
	$M \downarrow$.027	.027	.018	.021	.036	.027	.035	.028	.035	.027	.020
VT5000 [15]	$S \uparrow$.873	.867	.912	.895	.820	.839	.827	.839	.812	.843	.877
	$F_\beta \uparrow$.776	.801	.865	.846	.747	.801	.771	.769	.752	.817	.858
	$E_\xi \uparrow$.889	.897	.942	.927	.875	.891	.890	.883	.873	.903	.913
	$M \downarrow$.044	.043	.026	.033	.054	.048	.049	.048	.055	.042	.033
Ours vs. methods win/tie/loss		11/0/1	12/0/0	5/0/7	7/1/4	12/0/0	10/1/1	12/0/0	11/1/0	12/0/0	/	/

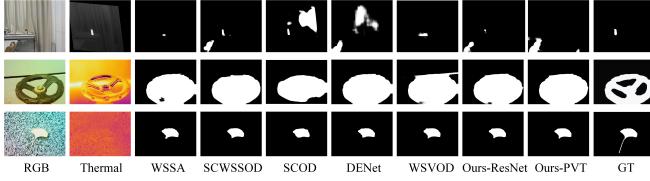


Fig. 5: Failure cases.

of scribble expansion via superpixels. Furthermore, when expanded labels are replaced with the predicted ones, the performance is improved. It discloses the fact that prediction module plays an important role in polishing the boundary noises of expanded labels and preventing the model from consistently learning false signals, generating the more accurate result.

Table 2: Ablation study about expanded labels and predicted labels.

Models	VT821 [16]			VT1000 [17]			VT5000 [15]					
	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$
Baseline	.873	.860	.921	.032	.914	.903	.937	.025	.858	.844	.914	.041
+Expanded	.889	.863	.928	.030	.919	.903	.941	.024	.873	.848	.926	.036
+Predicted	.895	.875	.934	.027	.925	.914	.949	.020	.877	.858	.931	.033

Multi-modal vs. Single-modal Ablation study about multi-modal supervision on two prediction branches is shown in Table 3. We can observe that multi-modal labels outperforms single-modal labels because the last two rows are better than the first two rows. Meanwhile, it is better to supervise each branch with the corresponding modality from the comparison between the last two rows. It is inferred from the data that the encoder can learn the local cue from superpixel segmentation via the corresponding multi-modal supervision.

Table 3: Ablation study about the supervision of two prediction modules. “X+Y”: X supervises the RGB branch and Y supervises the thermal branch, respectively.

Models	VT821 [16]			VT1000 [17]			VT5000 [15]					
	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$
RGB+RGB	.888	.866	.925	.028	.923	.915	.948	.021	.872	.848	.922	.035
Thermal+Thermal	.884	.869	.930	.029	.919	.910	.946	.023	.872	.849	.927	.035
Thermal+RGB	.890	.871	.935	.028	.922	.906	.944	.021	.877	.852	.928	.034
RGB+Thermal	.895	.875	.934	.027	.925	.914	.949	.020	.877	.858	.931	.033

Scribble vs. Aggregated Pseudo Label Ablation study about scribble supervision and aggregated pseudo label ones

is shown in Table 4. When our model removes the original scribble supervision, the performance is reduced from the comparison between the 1st and 3rd rows. It demonstrates that aggregated pseudo labels are inaccurate. So we need to continue to restrain the model via scribble supervision so that the model can better use the information provided by scribble annotation without being affected by the inaccurate pseudo labels. Of course, it is not enough to supervise the model only using the sparse scribble annotations, which can be seen from the comparison between the 2nd and 3rd rows. Scribble annotations which only utilize RGB information can't provide enough constraint when RGB images are not good. The additional pseudo label supervision can improve the performance.

Table 4: Ablation study about the supervision of scribble and aggregated pseudo labels.

Models	VT821 [16]			VT1000 [17]			VT5000 [15]					
	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$
w/o scribble label	.872	.820	.911	.034	.903	.879	.936	.028	.855	.789	.904	.040
w/o pseudo label	.873	.860	.921	.032	.914	.903	.937	.025	.858	.844	.914	.041
Ours	.895	.875	.934	.027	.925	.914	.949	.020	.877	.858	.931	.033

3.5. Model Generality Analysis

The model can be applied to RGB-D and video scribble-supervised SOD tasks.

In the experiments, RGB-D DENet [7] and video WSVOD [8] are retrained in both RGB-D and video scribble-supervised SOD datasets. In the top and bottom of Table 5, six RGB-D and six video SOD datasets are tested, respectively. From the result, we find our ResNet version wins 20 out of 24 compared with DENet in the RGB-D task and thoroughly defeat WSVOD. Meanwhile, in the video task, compared with WSVOD and DENet, our model wins out in most evaluation metrics. It is very impressive that our PVT version achieves the state-of-the-art performance. It verifies the generality of our model. It is first time to use a unified scribble-supervised model to run RGB-T, RGB-D, and video scribble-supervised SOD dataset. By the analysis, we find DENet and WSVOD only use multi-modal fusion in model structures and don't excavate the multi-modal cues in the supervision. Our model is

Table 5: Comparison with scribble-supervised multi-modal SOD methods in RGB-D and video datasets. **Red:** the best. **Blue:** the second best.

RGB-D SOD	NJU2K				STERE				DES				NLPR				LFSD				SIP			
	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$
DENet(TIP22) [7]	.875	.861	.902	.053	.874	.847	.910	.051	.898	.882	.954	.029	.898	.859	.944	.032	.827	.820	.866	.085	.846	.827	.905	.063
WSVOD(CVPR21) [8]	.814	.804	.861	.078	.835	.823	.887	.066	.844	.826	.908	.043	.863	.833	.927	.041	.751	.744	.782	.136	.759	.728	.849	.105
Ours-ResNet	.859	.860	.903	.058	.875	.873	.923	.047	.905	.902	.959	.023	.897	.879	.947	.029	.838	.849	.882	.076	.851	.846	.912	.059
Ours-PVT	.885	.886	.925	.044	.887	.876	.929	.042	.913	.897	.961	.023	.906	.884	.958	.024	.853	.862	.901	.068	.855	.851	.916	.056
Video SOD	SegV2				DAVSOD				VOS				FBMS				DAVIS				Visal			
	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$
WSVOD(CVPR21) [8]	.785	.685	.866	.032	.697	.564	.743	.101	.713	.607	.743	.085	.663	.650	.736	.064	.814	.712	.879	.037	.766	.706	.834	.036
DENet(TIP22) [7]	.746	.585	.807	.045	.675	.513	.712	.113	.692	.574	.715	.095	.652	.589	.706	.078	.788	.662	.844	.044	.746	.655	.807	.045
Ours-ResNet	.741	.725	.821	.028	.673	.587	.752	.088	.615	.570	.700	.096	.650	.670	.717	.070	.808	.789	.907	.033	.781	.793	.846	.024
Ours-PVT	.791	.792	.881	.023	.723	.657	.789	.077	.757	.722	.811	.056	.691	.715	.745	.048	.836	.827	.934	.027	.785	.798	.844	.022

the consistently excellent because we innovatively play the role of the other modality in the supervision.

4. CONCLUSION

In the paper, a scribble-supervised RGB-T SOD model is first proposed based on relabeled RGBT-S dataset. The sparse scribble annotation is expanded via superpixel segmentation in RGB and thermal images. Due to coarse boundary of the expanded label, a prediction module is used to predict the pseudo labels. The predicted multi-modal pseudo labels are aggregated to supervise the model training. The model is also applied to RGB-D and video SOD tasks and achieves the state-of-the-art performance. In the future, we will explore the scribble expansion solution via multi-modal supervoxels, in which the latent 3D geometry information will be further considered to improve the performance.

5. REFERENCES

- [1] Yao Wei and Shunping Ji, “Scribble-Based Weakly Supervised Deep Learning for Road Surface Extraction from Remote Sensing Images,” *TGRS*, vol. 60, pp. 1–12, 2021.
- [2] Xuan Li, Yuhang Xu, Lei Ma, Zhenghua Huang, and Haiwen Yuan, “Progressive Attention-Based Feature Recovery with Scribble Supervision for Saliency Detection in Optical Remote Sensing Image,” *TGRS*, vol. 60, pp. 1–12, 2022.
- [3] Xiangde Luo, Minhao Hu, Wenjun Liao, Shuwei Zhai, Tao Song, Guotai Wang, and Shaoting Zhang, “Scribble-Supervised Medical Image Segmentation via Dual-Branch Network and Dynamically Mixed Pseudo Labels Supervision,” *MICCAI*, pp. 528–538, 2022.
- [4] Qiuhibi Chen and Yi Hong, “Scribble2D5: Weakly-Supervised Volumetric Image Segmentation via Scribble Annotations,” *MICCAI*, pp. 1–10, 2022.
- [5] Ozan Unal, Dengxin Dai, and Luc Van Gool, “Scribble-Supervised Lidar Semantic Segmentation,” in *CVPR*, 2022, pp. 2697–2707.
- [6] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai, “Weakly-Supervised Salient Object Detection via Scribble Annotations,” in *CVPR*, 2020, pp. 12546–12555.
- [7] Yunqiu Xu, Xin Yu, Jing Zhang, Linchao Zhu, and Dadong Wang, “Weakly Supervised RGB-D Salient Object Detection With Prediction Consistency Training and Active Scribble Boosting,” *TIP*, vol. 31, pp. 2148–2161, 2022.
- [8] Wangbo Zhao, Jing Zhang, Long Li, Nick Barnes, Nian Liu, and Junwei Han, “Weakly Supervised Video Salient Object Detection,” in *CVPR*, 2021, pp. 16826–16835.
- [9] Shuyong Gao, Haozhe Xing, Wei Zhang, Yan Wang, Qianyu Guo, and Wenqiang Zhang, “Weakly Supervised Video Salient Object Detection via Point Supervision,” in *ACM MM*, 2022, pp. 3656–3665.
- [10] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk, “SLIC Superpixels Compared to State-of-the-art Superpixel Methods,” *PAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.
- [11] Siyu Yu, Bingfeng Zhang, Jimin Xiao, and Eng Gee Lim, “Structure-consistent weakly supervised salient object detection with local saliency coherence,” in *AAAI*, 2021, pp. 3234–3242.
- [12] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du, “Learning Affinity from Attention: End-to-End Weakly-Supervised Semantic Segmentation with Transformers,” in *CVPR*, 2022, pp. 16846–16855.
- [13] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz, “Pixel-Adaptive Convolutional Neural Networks,” in *CVPR*, 2019, pp. 11166–11175.
- [14] Jun Wei, Shuhui Wang, and Qingming Huang, “F³Net: Fusion, Feedback and Focus for Salient Object Detection,” in *AAAI*, 2020, vol. 34, pp. 12321–12328.
- [15] Zhengzheng Tu, Yan Ma, Zhun Li, Chenglong Li, Jieming Xu, and Yongtao Liu, “RGB-T Salient Object Detection: A Large-Scale Dataset and Benchmark,” *TMM*, pp. 1–14, 2022.
- [16] Guizhao Wang, Chenglong Li, Yunpeng Ma, Aihua Zheng, Jin Tang, and Bin Luo, “RGB-T Saliency Detection Benchmark: Dataset, Baselines, Analysis and a Novel Approach,” in *CCIG*. Springer, 2018, pp. 359–369.
- [17] Zhengzheng Tu, Tian Xia, Chenglong Li, Xiaoxiao Wang, Yan Ma, and Jin Tang, “RGB-T Image Saliency Detection via Collaborative Graph Learning,” *TMM*, vol. 22, no. 1, pp. 160–173, 2019.
- [18] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao, “BBS-Net: RGB-D Salient Object Detection with a Bifurcated Backbone Strategy Network,” in *ECCV*. Springer, 2020, pp. 275–292.
- [19] Zhengyi Liu, Yacheng Tan, Qian He, and Yun Xiao, “SwinNet: Swin Transformer Drives Edge-Aware RGB-D and RGB-T Salient Object Detection,” *TCSV*, vol. 32, pp. 4486–4497, 2021.
- [20] Diederik P Kingma and Jimmy Ba, “Adam: A Method for Stochastic Optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [21] Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang, “Multi-Interactive Dual-Decoder for RGB-Thermal Salient Object Detection,” *TIP*, vol. 30, pp. 5678–5691, 2021.
- [22] Runmin Cong, Kepu Zhang, Chen Zhang, Feng Zheng, Yao Zhao, Qingming Huang, and Sam Kwong, “Does Thermal Really Always Matter for RGB-T Salient Object Detection?,” *TMM*, pp. 1–12, 2022.
- [23] Ruozhen He, Qihua Dong, Jiaying Lin, and Rynson WH Lau, “Weakly-Supervised Camouflaged Object Detection with Scribble Annotations,” in *AAAI*, 2023, pp. 1–12.
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep Residual Learning for Image Recognition,” in *CVPR*, 2016, pp. 770–778.
- [25] Wenhui Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao, “Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions,” in *CVPR*, 2021, pp. 568–578.