IJCAI
JEJU 2024

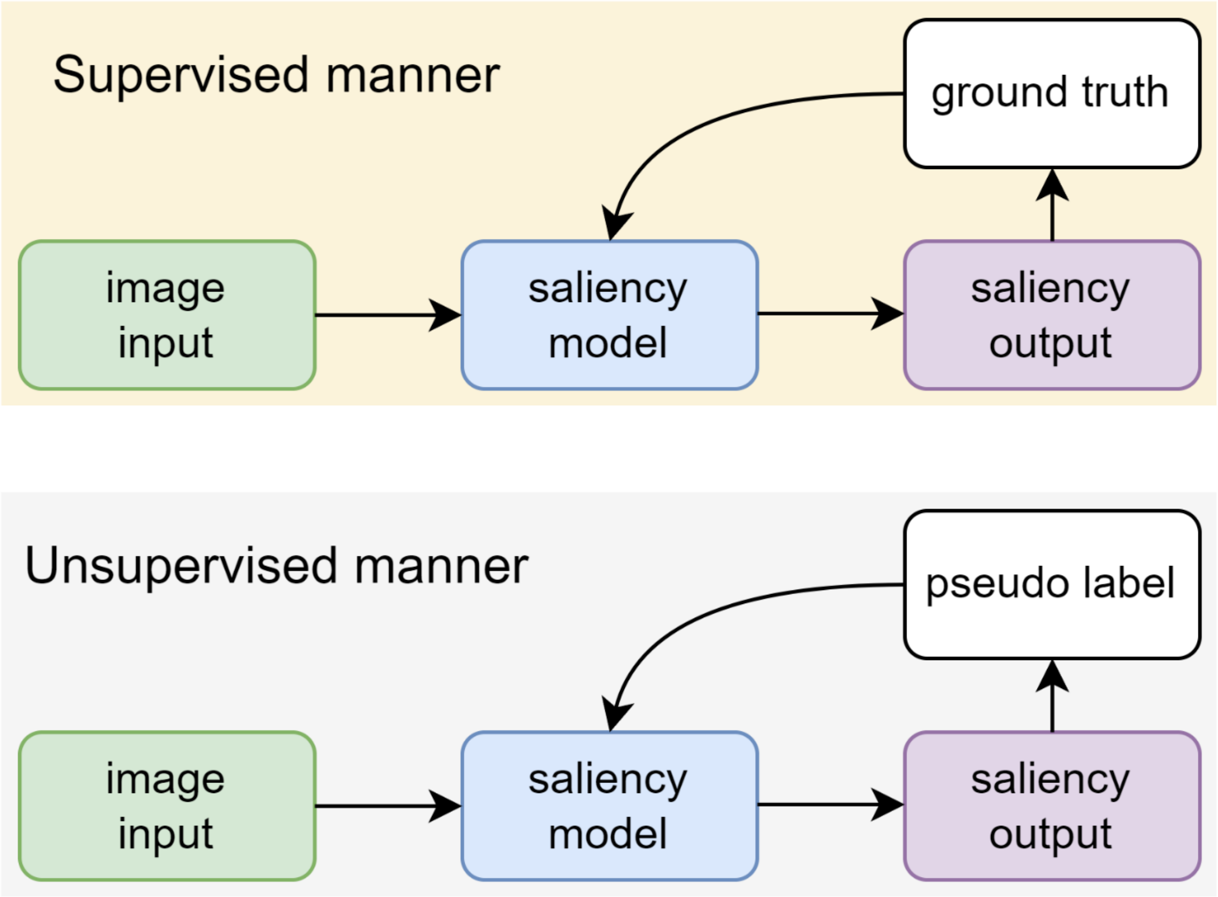
Unified Unsupervised Salient Object Detection via Knowledge Transfer

Yao Yuan , **Wutao Liu** , **Pan Gao*** , **Qun Dai** and **Jie Qin***
Nanjing University of Aeronautics and Astronautics
{ayews233, wutaoliu, pan.gao, daiqun}@nuaa.edu.cn, qinjiebuaa@gmail.com

01 Introduction

Salient object detection (SOD) aims to identify the most visually significant objects in images. Supervised SOD methods have achieved excellent results, but due to their heavy reliance on pixel-level annotations for salient objects, unsupervised SOD (USOD) has been gaining increasing attention.

USOD not only eliminates the need for annotated data but also exhibits strong generalization performance.



01 Introduction

However, current Unsupervised Salient Object Detection (USOD) methods still face the following three major challenges:

1. Obtaining initial saliency cues (pseudo-labels). Existing hand-crafted feature-based methods and deep learning methods still unable to complete this task well.

2. Enhancing pseudo-label quality. The performance of the saliency model heavily relies on the quality of pseudo-labels. Enhancing the quality of these labels can significantly improve the model's effectiveness.

3. Very limited exploration in migration. Current USOD methods predominantly target specific modalities like RGB, RGB-D, and RGB-T, yet they have not extensively investigated the potential benefits of unsupervised approaches in transfer learning and generalization.

01 Introduction

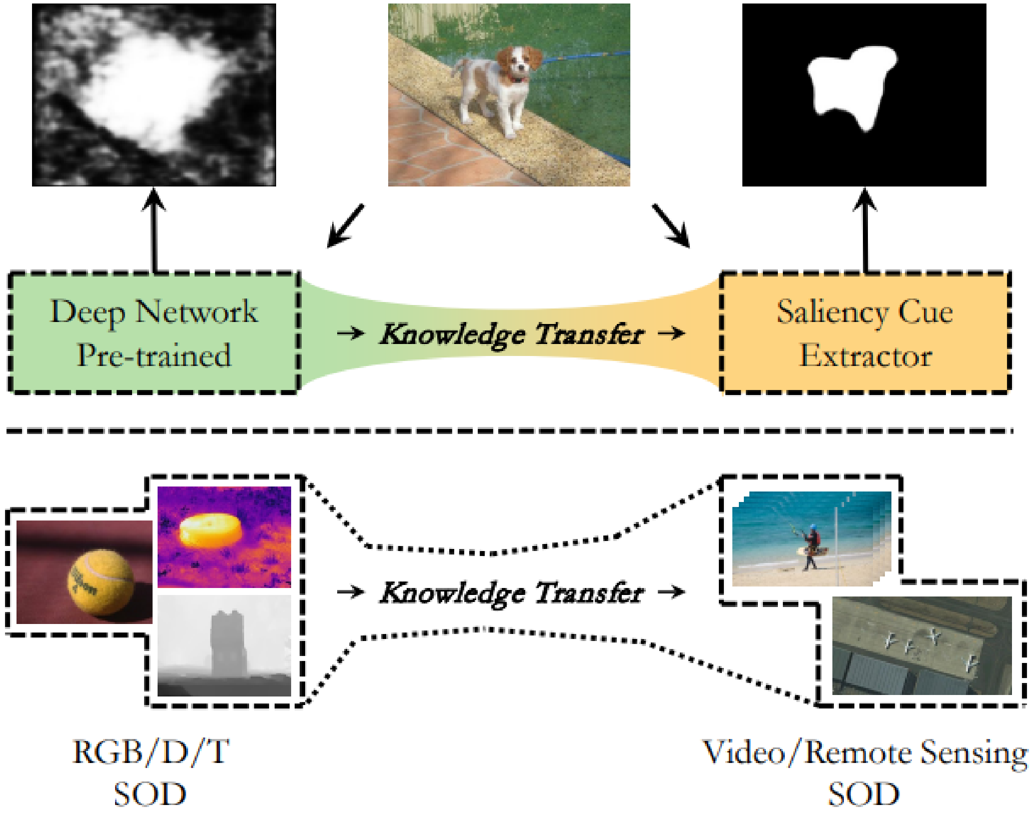


Figure 1: The proposed framework includes two types of knowledge transfer: (1) From pre-trained deep network to saliency cue extractor; (2) From Natural Still Image (NSI) SOD to non-NSI SOD.

Our method is grounded in knowledge transfer.

It incorporates two types of transfer: one extracts saliency cues from pre-trained deep networks (e.g., MoCo-v2), while the other transfers knowledge from Natural Still Image (NSI) SOD (including RGB, RGB-D, RGB-T) to non-NSI SOD.

Our contribution is fundamentally based on how to achieve these two types of knowledge transfer.

01 Introduction

Main contributions can be summarized as:

1) We propose the Progressive Curriculum Learning-based Saliency Distilling (PCL-SD) mechanism to extract saliency cues from easy samples to hard ones.

2) We design the Self-rectify Pseudo-label Refinement (SPR) mechanism to gradually improve the quality of pseudo-labels during the training process.

3) We devise an adapter-tuning method to transfer saliency knowledge from NSI SOD to non-NSI SOD tasks, achieving impressive transfer performance.

02 Methodology: Framework

We propose a two-stage framework for unified USOD tasks. In stage 1, we train a saliency cue extractor (SCE) to transfer saliency knowledge from a pre-trained deep network. In stage 2, we utilize the obtained saliency cues as initial pseudo-labels to train a saliency detector (SD). We train our base model on Natural Still Image datasets and subsequently transfer the model to non-NSI SOD tasks such as video SOD and RSI SOD.

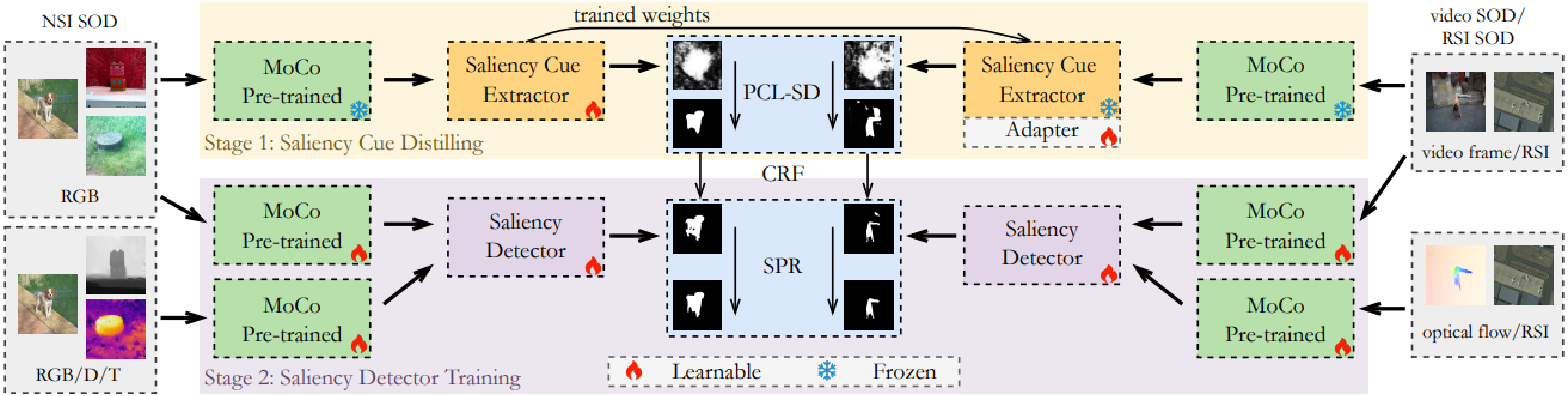


Figure 2: Overview of the proposed method. The left side represents the training process on NSI SOD, while the right side shows the training process of transferring to non-NSI SOD tasks.

02 Methodology: Progressive Curriculum Learning-based Saliency Distilling

The PCL-SD mechanism rigidly excludes hard samples at the early stages of training and gradually incorporates them as training progresses. As a result, the model progressively extracts saliency knowledge from easy to hard samples, and the entire training process becomes more robust and stable.

$$\mathcal{L}_{sal} = 0.5 - \frac{1}{N}\sum_{i}^{N} \|S(i) - 0.5|$$

The basic saliency distilling.

$$|S(i) - 0.5| < p.$$

The partitioning of hard samples.

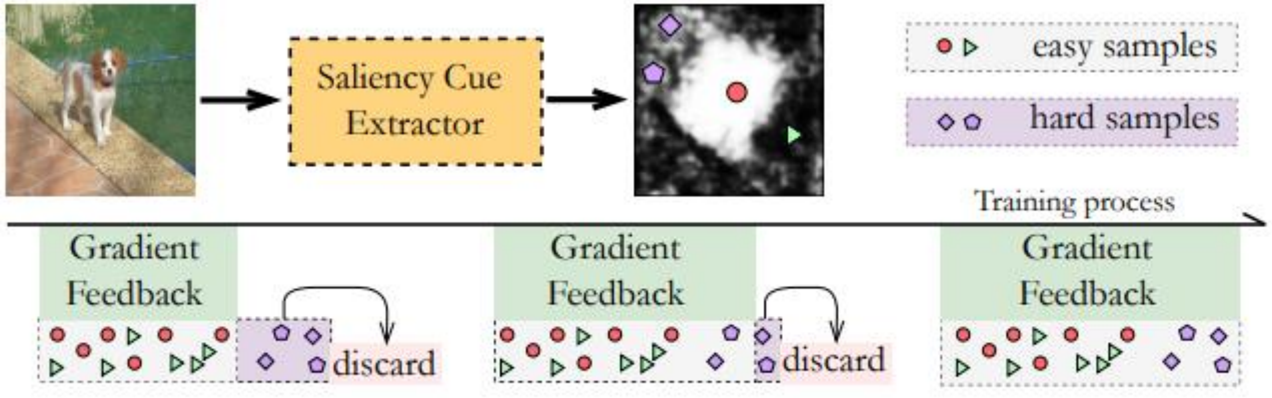


Figure 3: Illustration of the proposed PCL-SD. Hard samples are progressively incorporated as the training progresses.

$$M(i) = \begin{cases} 0 & if\ |S(i) - 0.5| < p, \\ 1 & otherwise, \end{cases}$$

$$\mathcal{L}_{pcl-sd} = 0.5 - \frac{1}{N}\sum_{i}^{N} |M(i) \odot S(i) - 0.5|$$

The proposed PCL-SD.

02 Methodology: Self-rectify Pseudo-label Refinement (SPR)

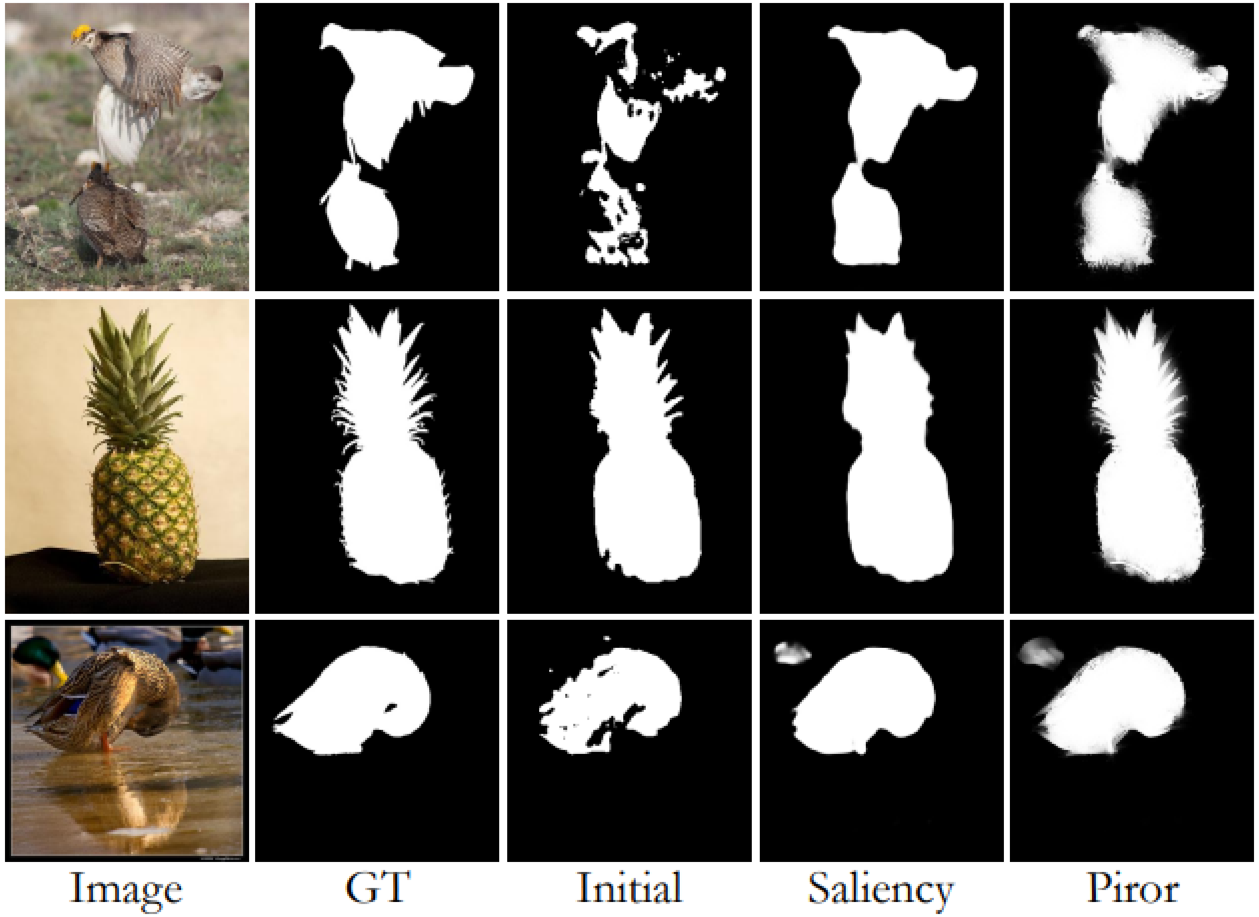


Figure 4: The comparison between initial pseudo-label, saliency prediction, and prior rectification.

From the left image, it is evident that the initial pseudo-labels, the posterior rectifications from the model's saliency prediction, and the prior rectification from a real-time pixel refiner each possess distinct advantages and disadvantages.

The SPR mechanism effectively integrates prior and posterior rectifications, gradually improving the quality of pseudo labels during the training process, demonstrating strong self-supervised performance.

$$G_{\text{ref}} = \lambda_1 R_{\text{pri}} + \lambda_2 R_{\text{post}} + \lambda_3 G_{\text{pre}}$$

The process of updating pseudo labels.

02 Methodology: Knowledge Transfer via Adapter-tuning

We devise an adapter-tuning method to transfer the acquired saliency knowledge, leveraging shared knowledge to attain superior transferring performance on the target tasks. Specifically, we employed an Adapter structure to fine-tune the deep layers of the model, allowing it to adapt to the target task without experiencing performance degradation.

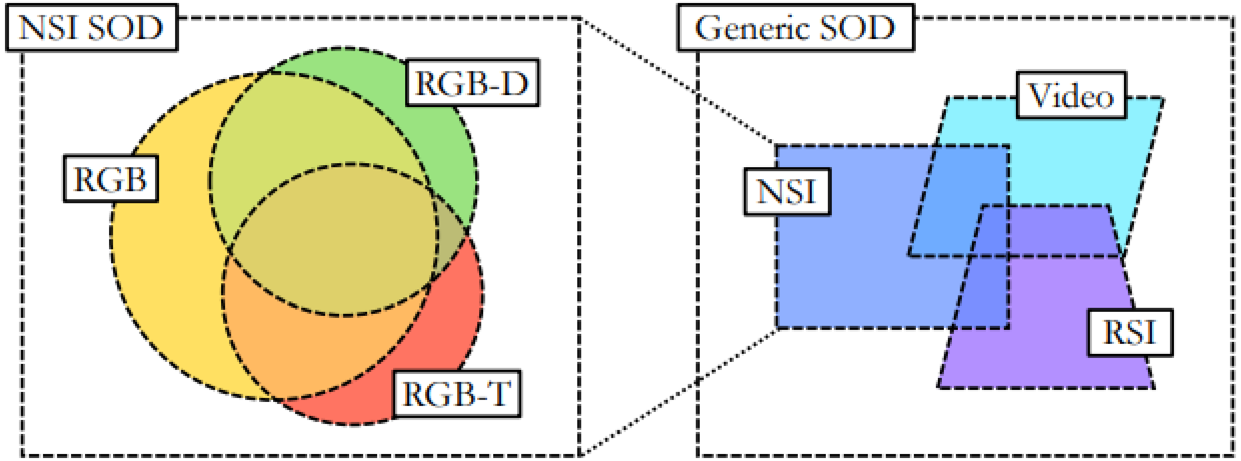


Figure 5: The relevance between different SOD tasks. The overlaps can be seen as shared common knowledge.

$$\hat{F} = T(F)$$

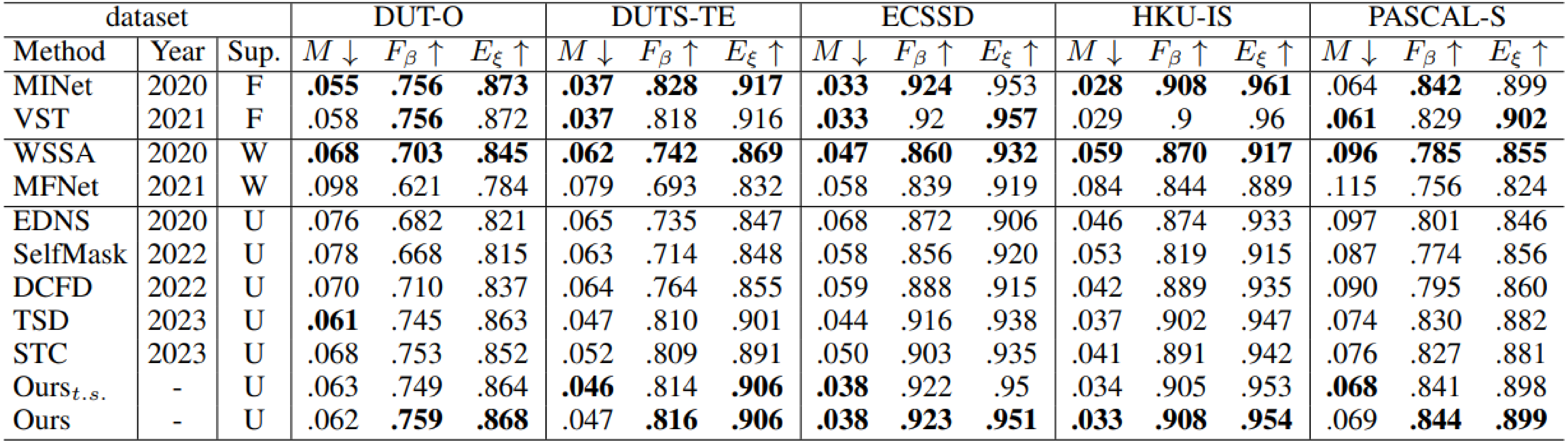
w/o adapter

$$\hat{F} = T(F) + T_a(F)$$

w/ adapter

03 Experiments: Experimental Results

Quantitative Comparison:

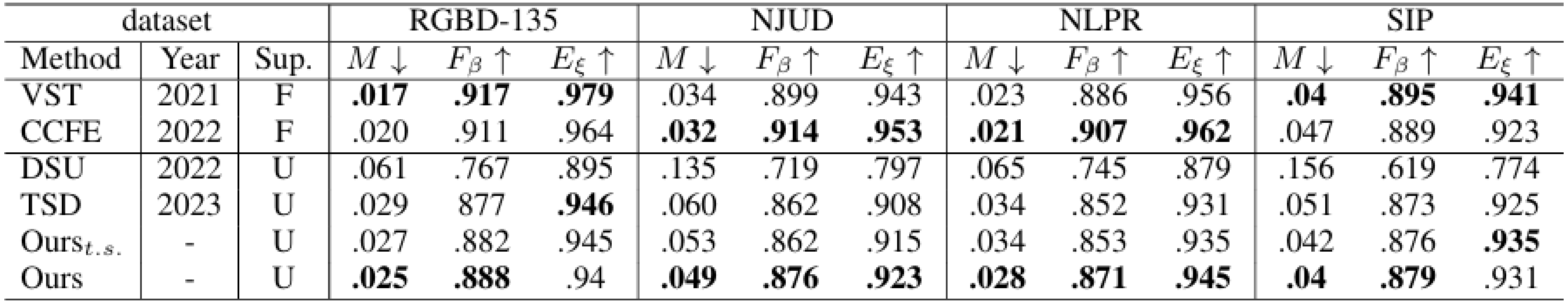


dataset			DUT-O			DUTS-TE			ECSSD			HKU-IS			PASCAL-S		
Method	Year	Sup.	$M \downarrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\xi \uparrow$
MINet	2020	F	**.055**	**.756**	**.873**	**.037**	**.828**	**.917**	**.033**	**.924**	.953	**.028**	**.908**	**.961**	.064	**.842**	.899
VST	2021	F	.058	**.756**	.872	**.037**	.818	.916	**.033**	.92	**.957**	.029	.9	.96	**.061**	.829	**.902**
WSSA	2020	W	**.068**	**.703**	**.845**	**.062**	**.742**	**.869**	**.047**	**.860**	**.932**	**.059**	**.870**	**.917**	**.096**	**.785**	**.855**
MFNet	2021	W	.098	.621	.784	.079	.693	.832	.058	.839	.919	.084	.844	.889	.115	.756	.824
EDNS	2020	U	.076	.682	.821	.065	.735	.847	.068	.872	.906	.046	.874	.933	.097	.801	.846
SelfMask	2022	U	.078	.668	.815	.063	.714	.848	.058	.856	.920	.053	.819	.915	.087	.774	.856
DCFD	2022	U	.070	.710	.837	.064	.764	.855	.059	.888	.915	.042	.889	.935	.090	.795	.860
TSD	2023	U	**.061**	.745	.863	.047	.810	.901	.044	.916	.938	.037	.902	.947	.074	.830	.882
STC	2023	U	.068	.753	.852	.052	.809	.891	.050	.903	.935	.041	.891	.942	.076	.827	.881
Ours$_{t.s.}$	-	U	.063	.749	.864	**.046**	.814	**.906**	**.038**	.922	.95	.034	.905	.953	**.068**	.841	.898
Ours	-	U	.062	**.759**	**.868**	.047	**.816**	**.906**	**.038**	**.923**	**.951**	**.033**	**.908**	**.954**	.069	**.844**	**.899**

Table 1: Quantitative comparison on RGB SOD benchmarks. “Sup.” indicates the supervised signals used to train SOD methods. “F”, “W” and “U” mean fully-supervised, weakly-supervised and unsupervised, respectively. The best results are shown in **bold**.

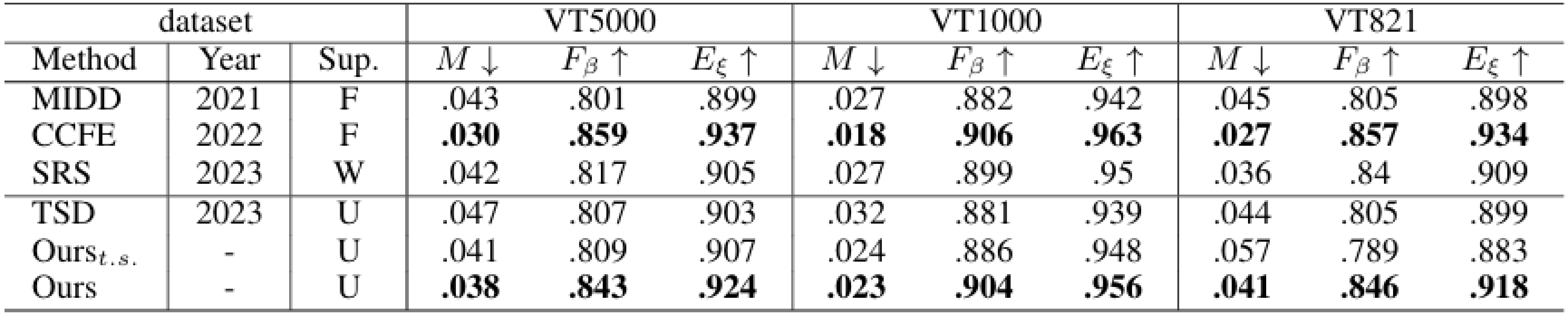
03 Experiments: Experimental Results

Quantitative Comparison:



dataset			RGBD-135			NJUD			NLPR			SIP		
Method	Year	Sup.	$M \downarrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\xi \uparrow$
VST	2021	F	**.017**	**.917**	**.979**	.034	.899	.943	.023	.886	.956	**.04**	**.895**	**.941**
CCFE	2022	F	.020	.911	.964	**.032**	**.914**	**.953**	**.021**	**.907**	**.962**	.047	.889	.923
DSU	2022	U	.061	.767	.895	.135	.719	.797	.065	.745	.879	.156	.619	.774
TSD	2023	U	.029	877	**.946**	.060	.862	.908	.034	.852	.931	.051	.873	.925
$\text{Ours}_{t.s.}$	-	U	.027	.882	.945	.053	.862	.915	.034	.853	.935	.042	.876	**.935**
Ours	-	U	**.025**	**.888**	.94	**.049**	**.876**	**.923**	**.028**	**.871**	**.945**	**.04**	**.879**	.931

Table 2: Quantitative comparison on RGB-D SOD benchmarks.

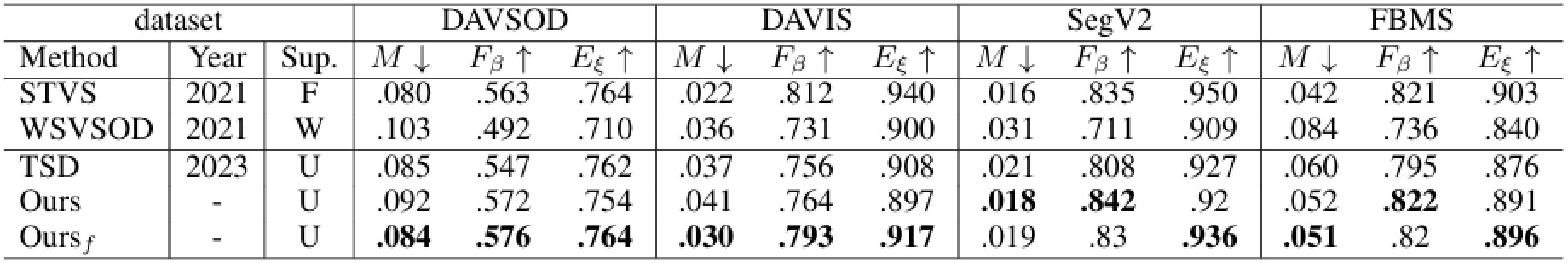


dataset			VT5000			VT1000			VT821		
Method	Year	Sup.	$M \downarrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\xi \uparrow$
MIDD	2021	F	.043	.801	.899	.027	.882	.942	.045	.805	.898
CCFE	2022	F	**.030**	**.859**	**.937**	**.018**	**.906**	**.963**	**.027**	**.857**	**.934**
SRS	2023	W	.042	.817	.905	.027	.899	.95	.036	.84	.909
TSD	2023	U	.047	.807	.903	.032	.881	.939	.044	.805	.899
$\text{Ours}_{t.s.}$	-	U	.041	.809	.907	.024	.886	.948	.057	.789	.883
Ours	-	U	**.038**	**.843**	**.924**	**.023**	**.904**	**.956**	**.041**	**.846**	**.918**

Table 3: Quantitative comparison on RGB-T SOD benchmarks.

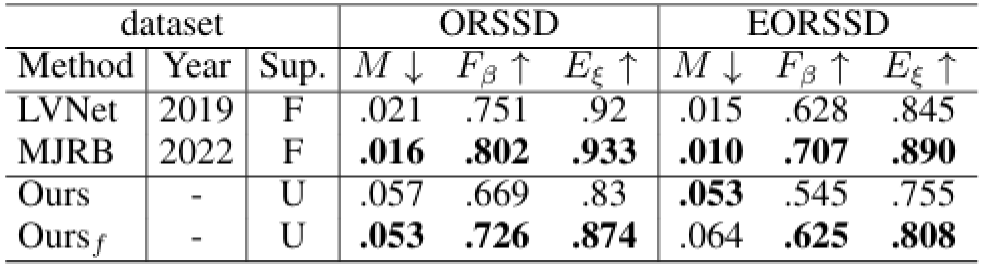
03 Experiments: Experimental Results

Quantitative Comparison:



dataset			DAVSOD			DAVIS			SegV2			FBMS		
Method	Year	Sup.	$M\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$
STVS	2021	F	.080	.563	.764	.022	.812	.940	.016	.835	.950	.042	.821	.903
WSVSOD	2021	W	.103	.492	.710	.036	.731	.900	.031	.711	.909	.084	.736	.840
TSD	2023	U	.085	.547	.762	.037	.756	.908	.021	.808	.927	.060	.795	.876
Ours	-	U	.092	.572	.754	.041	.764	.897	**.018**	**.842**	.92	.052	**.822**	.891
Ours$_f$	-	U	**.084**	**.576**	**.764**	**.030**	**.793**	**.917**	.019	.83	**.936**	**.051**	.82	**.896**

Table 4: Quantitative comparison on video SOD benchmarks.



dataset			ORSSD			EORSSD		
Method	Year	Sup.	$M\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$
LVNet	2019	F	.021	.751	.92	.015	.628	.845
MJRB	2022	F	**.016**	**.802**	**.933**	**.010**	**.707**	**.890**
Ours	-	U	.057	.669	.83	**.053**	.545	.755
Ours$_f$	-	U	**.053**	**.726**	**.874**	.064	**.625**	**.808**

Table 5: Quantitative comparison on RSI SOD benchmarks.

03 Experiments: Experimental Results

Qualitative Comparison:

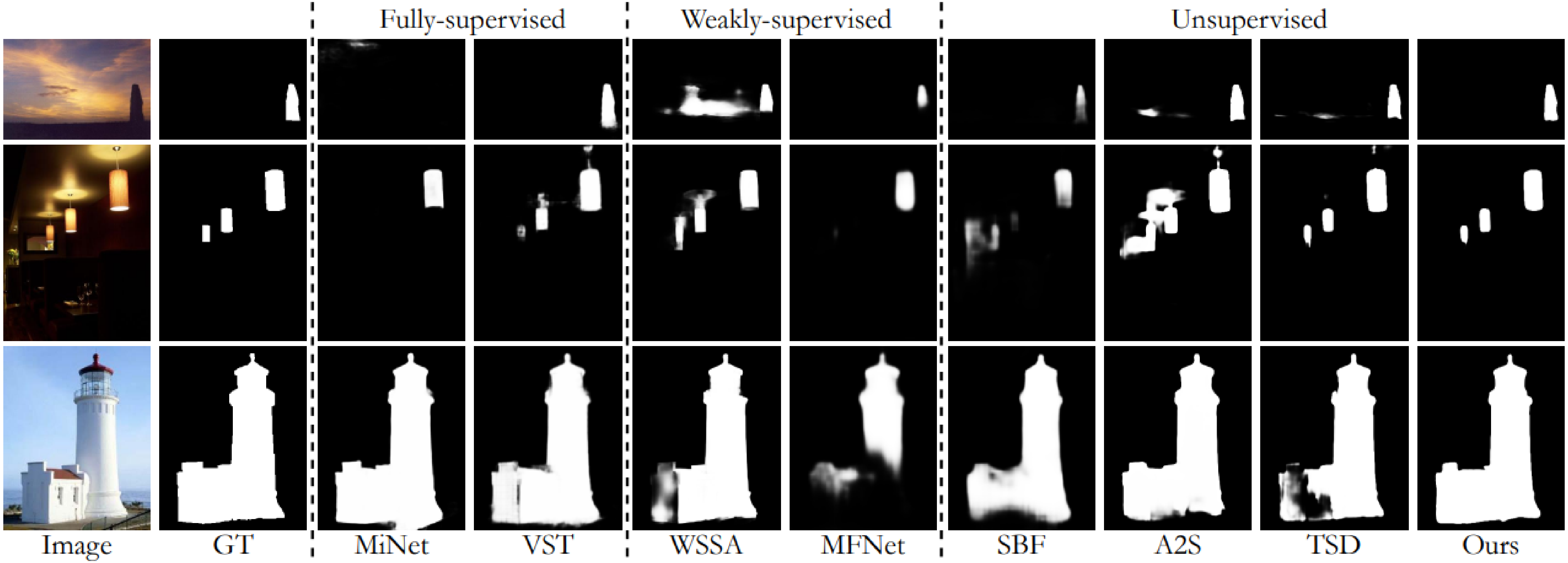


Figure 6: Visual comparison between the proposed method and other state-of-the-art SOD methods on RGB SOD datasets.

Experiments: Experimental Results

Qualitative Comparison:

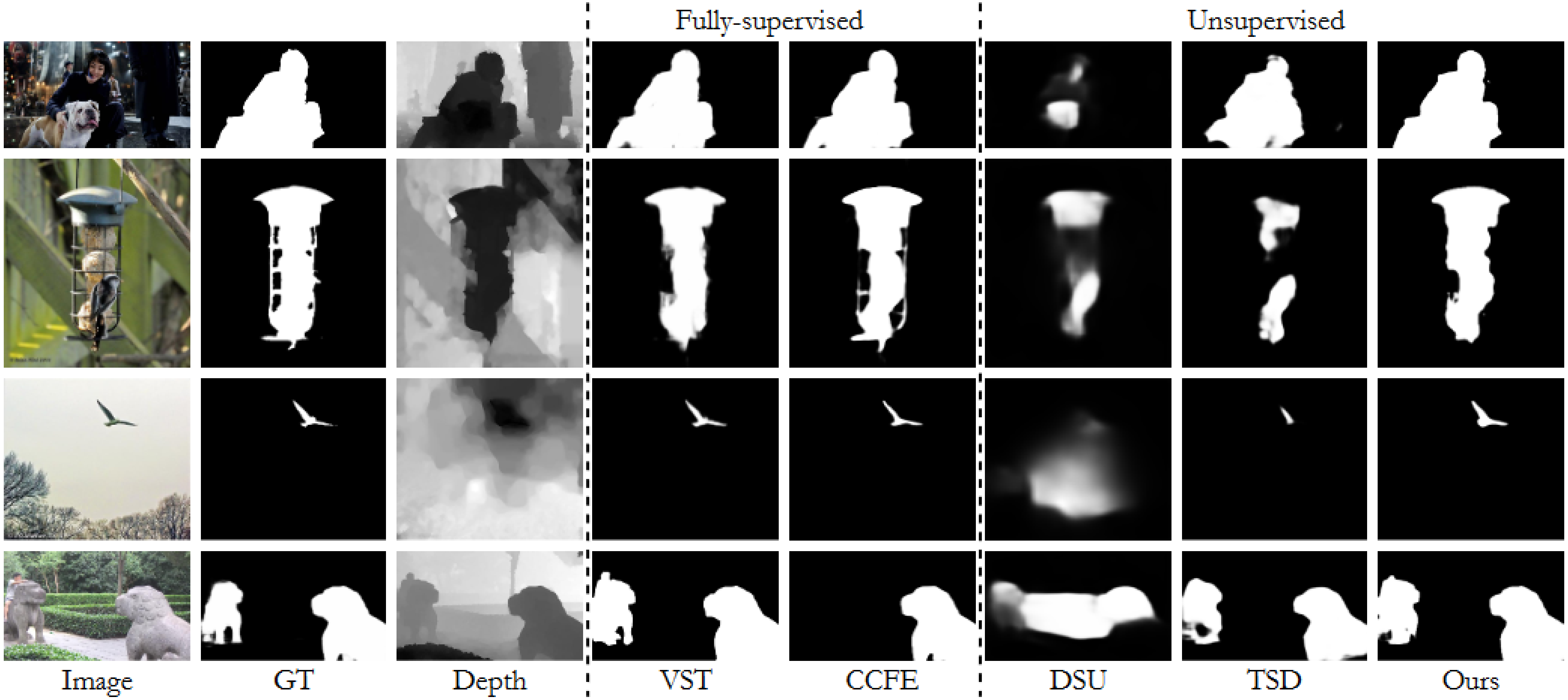


Figure 10: Visual comparison between the proposed method and the other state-of-the-art SOD methods on RGB-D SOD datasets.

03 Experiments: Experimental Results

Qualitative Comparison:

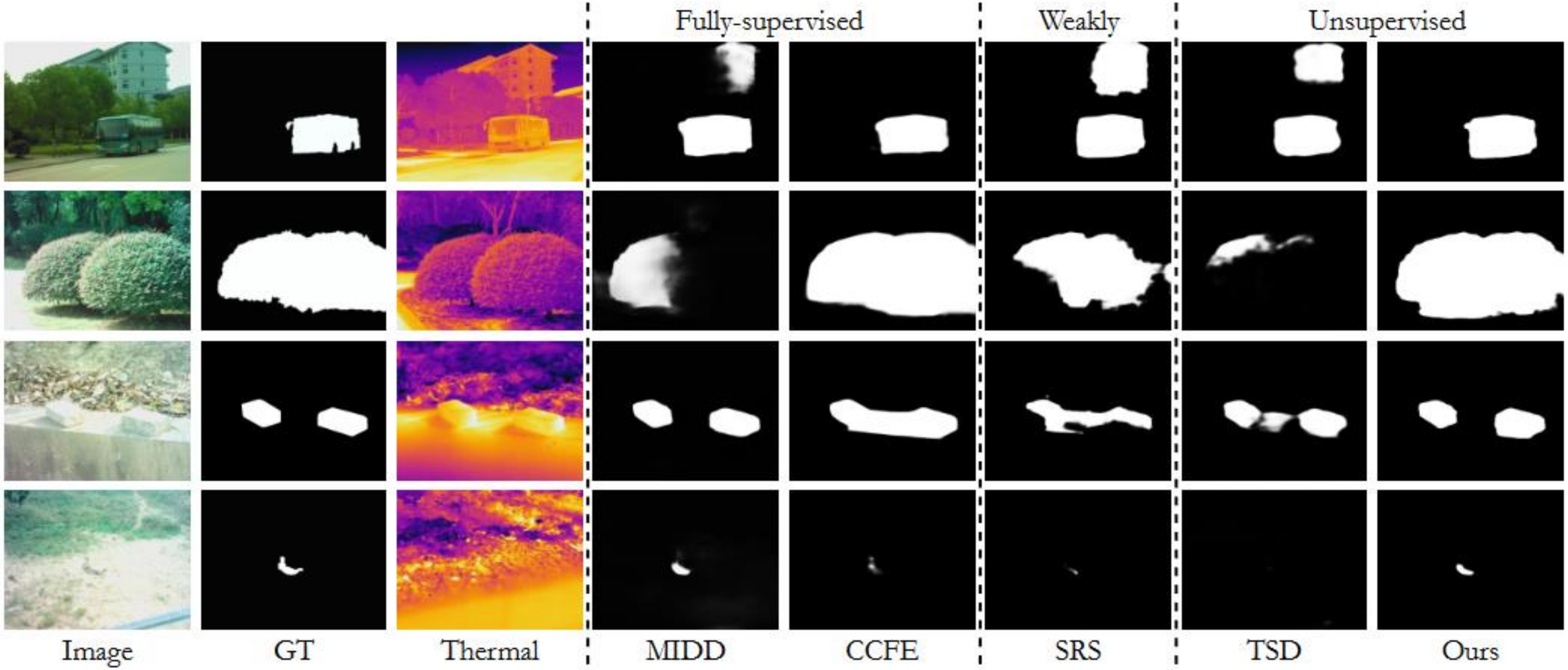


Figure 11: Visual comparison between the proposed method and the other state-of-the-art SOD methods on RGB-T SOD datasets.

03 Experiments: Experimental Results

Qualitative Comparison:

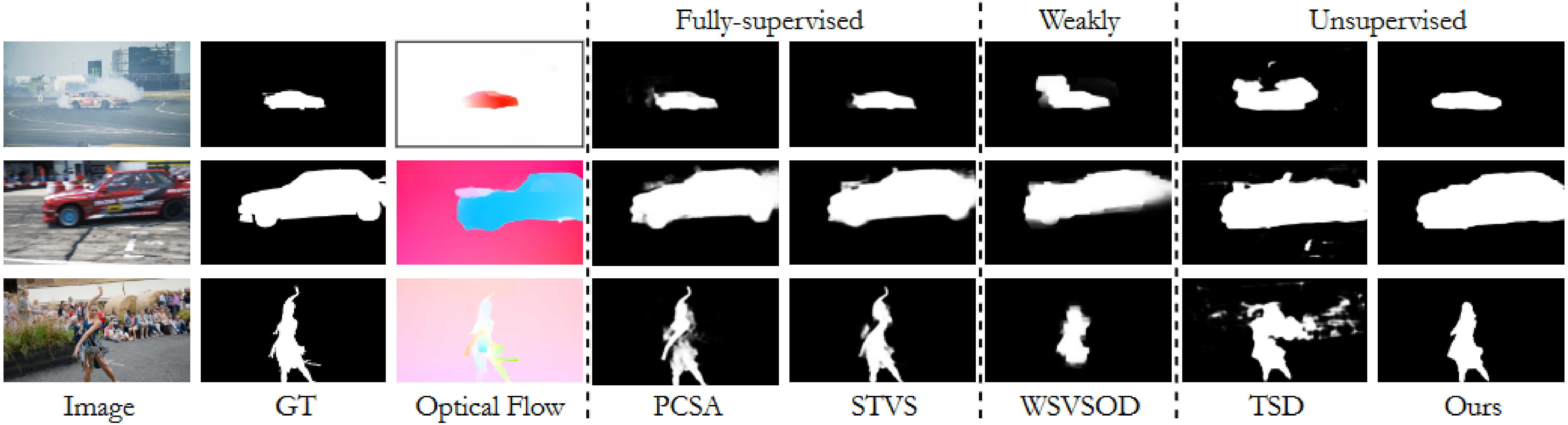


Figure 12: Visual comparison between the proposed method and the other state-of-the-art SOD methods on video SOD datasets.

03 Experiments: Experimental Results

Qualitative Comparison:

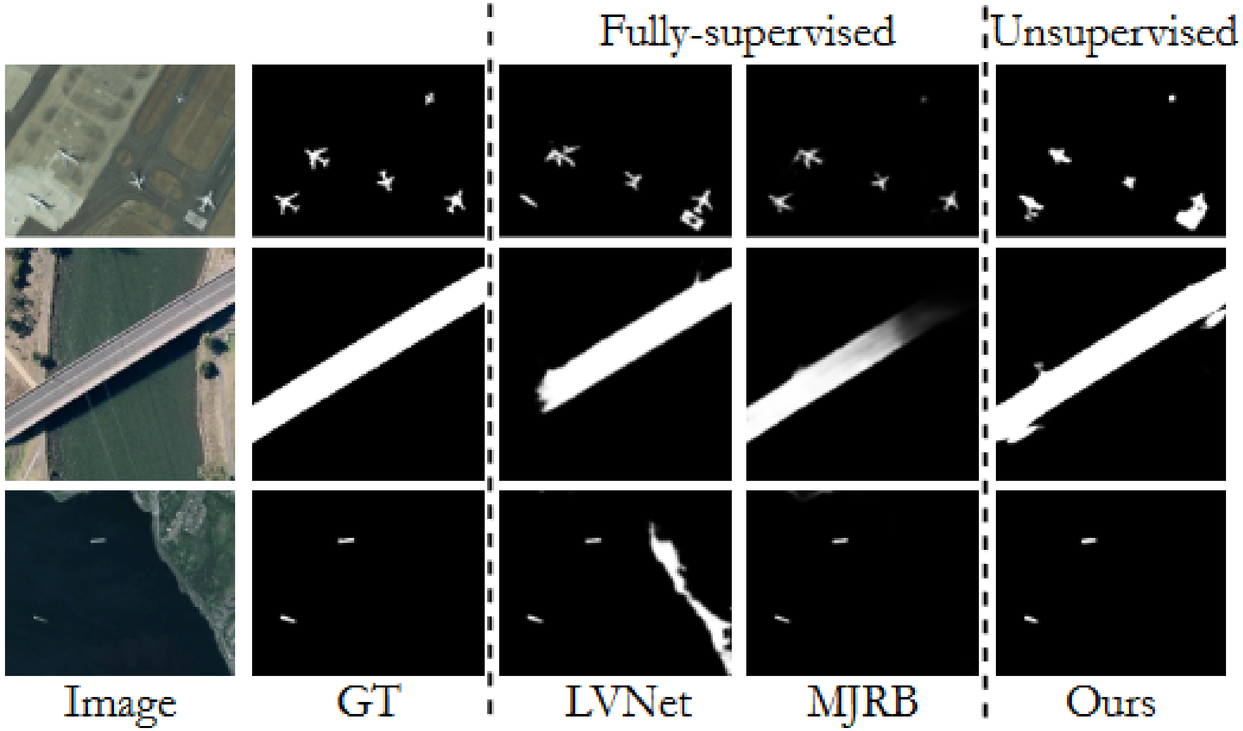
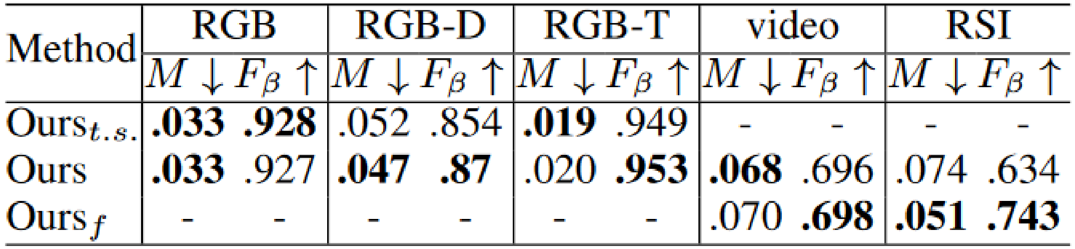


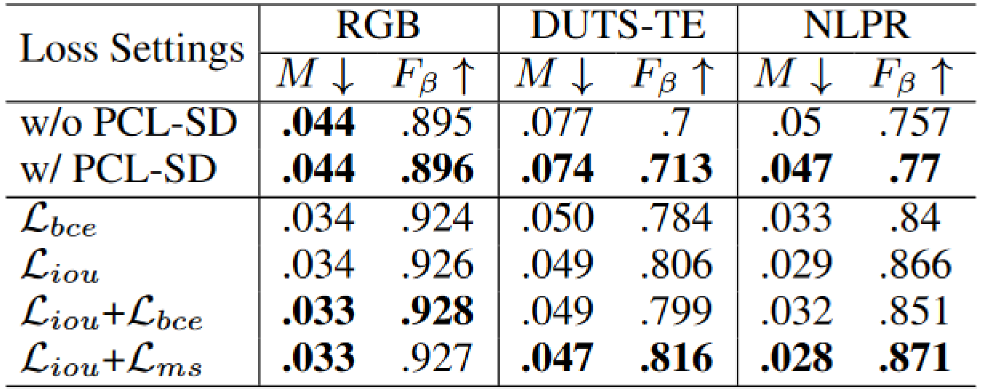
Figure 13: Visual comparison between the proposed method and the other state-of-the-art SOD methods on RSI SOD datasets.

03 Experiments: Ablation Study



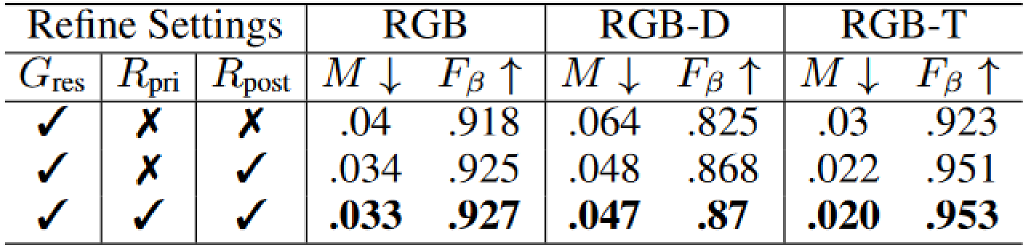
Method	RGB		RGB-D		RGB-T		video		RSI	
	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$
Ours$_{t.s.}$	**.033**	**.928**	.052	.854	**.019**	.949	-	-	-	-
Ours	**.033**	.927	**.047**	**.87**	.020	**.953**	**.068**	.696	.074	.634
Ours$_f$	-	-	-	-	-	-	.070	**.698**	**.051**	**.743**

Table 6: Evaluation on Pseudo-label Quality.



Loss Settings	RGB		DUTS-TE		NLPR	
	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$
w/o PCL-SD	**.044**	.895	.077	.7	.05	.757
w/ PCL-SD	**.044**	**.896**	**.074**	**.713**	**.047**	**.77**
$\mathcal{L}_{bce}$	.034	.924	.050	.784	.033	.84
$\mathcal{L}_{iou}$	.034	.926	.049	.806	.029	.866
$\mathcal{L}_{iou}+\mathcal{L}_{bce}$	**.033**	**.928**	.049	.799	.032	.851
$\mathcal{L}_{iou}+\mathcal{L}_{ms}$	**.033**	.927	**.047**	**.816**	**.028**	**.871**

Table 8: Evaluation on Supervision Strategy. “RGB” denotes the training set of RGB SOD.



Refine Settings			RGB		RGB-D		RGB-T	
G_{res}	R_{pri}	R_{post}	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$
✓	✗	✗	.04	.918	.064	.825	.03	.923
✓	✗	✓	.034	.925	.048	.868	.022	.951
✓	✓	✓	**.033**	**.927**	**.047**	**.87**	**.020**	**.953**

Table 7: Evaluation on Self-rectify Pseudo-label Refinement.

04 Conclusion&Discussion

This paper proposes a two-stage unified unsupervised SOD framework for generic SOD tasks, with knowledge transfer as the foundation.

Merits:

- More stable distilling of saliency cues.
- Improved pseudo-label refinement.
- Adapter-tuning driven knowledge transfer.

Limitations:

- From Modality-Agnostic to Modality-Informed: During the training process, the model treats different modalities, such as depth, thermal, and optical flow, in a modality-agnostic manner, which may limit the model’s ability to exploit saliency information specific to each modality.

- Deeper and More Targeted Migration: We treated both tasks as non-NSI SOD tasks without incorporating more targeted adaptation measures. For instance, in video SOD, we did not directly input video data into the model but treated each frame as a separate two-dimensional image.

Thank you for listening!

For more information, please check:

Github

WeChat

I2ML Website