Unified Unsupervised Salient Object Detection via Knowledge Transfer

Supplementary Material

1 Motivation

Unsupervised salient object detection (USOD) has emerged as a solution to address the reliance on large-scale annotated data, enabling models to train on more extensive datasets. The annotation-free nature of unsupervised SOD facilitates stronger generalization performance. However, in the case of certain fundamental SOD datasets (e.g., MSRA [Liu *et al.*, 2007], DUTS [Wang *et al.*, 2017], VT5000 [Tu *et al.*, 2022a]), substantial annotated data already exists, and existing supervised methods have achieved good performance. When it comes to basic Natural Still Image (NSI) SOD, such as RGB, RGB-D, and RGB-T tasks, unsupervised methods can only avoid overfitting on annotations but still exhibit significant performance gaps compared to supervised methods. However, for more advanced SOD tasks, such as video SOD and remote sense image (RSI) SOD, the challenges lie not only in annotation difficulties but also in high data collection thresholds or costs. Limited datasets in these scenarios make it prone to overfitting and overconfidence during training. Additionally, the limited amount of data makes training a model from scratch without supervision challenging to achieve desirable performance. Currently, no unsupervised methods exist for RSI-SOD, and research on unsupervised SOD in these tasks has been largely neglected.

We posit that there is a correlation between different SOD tasks, implying the presence of shared common knowledge. Existing studies have attempted to develop a universal approach by incorporating multiple SOD tasks into a unified framework. Specifically, [Wang *et al.*, 2023] attempted prompt tuning to address the inconsistency between modalities. However, these investigations are largely confined to the NSI domain. As we broaden our perspective to encompass a wider range of SOD domains, the shared knowledge between tasks becomes more limited, and the differences become the primary influencing factors. In this paper, we introduce video SOD and RSI SOD tasks to establish the notion of generic SOD, which can also include other specific tasks like hyperspectral salient object detection, underwater salient object detection, and so on. Therefore, training a unified model by combining data from all tasks to address the generic SOD task is currently impractical.

Despite the increase in differences, we firmly hold the belief that common saliency knowledge exists. Specifically, we contend that video SOD and RSI SOD tasks can learn valuable knowledge from the NSI domain. Unsupervised approaches, unlike supervised methods, are not hindered by manual annotations, enabling them to focus on this shared common knowledge and exhibit better generalization and transferability. Previous research [Zhou *et al.*, 2023b] has further demonstrated that unsupervised methods achieve superior zero-shot transfer results compared to existing supervised methods in scenarios such as chest X-ray images. Based on the aforementioned analysis, we assert that the exploration of unsupervised SOD methods for generic SOD tasks is both valuable and meaningful. It fosters research in data-scarce SOD tasks and propels SOD advancements in real-world scenarios.

2 Details of the proposed framework

Our proposed framework employs a saliency cue extractor (SCE) and a saliency detector (SD). The architecture of the SCE is illustrated in Figure 1. We utilize A2S [Zhou *et al.*, 2023a] as the underlying model and incorporate an adapter component for transfer learning. The SCE consists of ResNet as the backbone, along with several SE [Hu *et al.*, 2018] modules. Its objective is to obtain a multi-scale activation map from different layers of the backbone. The structure of SCE is similar to existing end-to-end SOD models, but deliberately kept simple to control network depth and obtain activation map outputs closer to the backbone. Whether training from scratch or fine-tuning, we maintain weight freezing of the backbone. This is because unsupervised training is susceptible to instability, and freezing the weights helps mitigate the risk of model degradation. During fine-tuning for transfer learning to other tasks, we assume that deep features are more task-specific, while shallow features are generally more general. Based on this assumption, we apply adapter tuning only to the deepest layer's features. Specifically, we add an additional SE module and use residual connections to add its output to the original network. Only the parameters of this module are involved in gradient backpropagation, while the remaining parameters are kept frozen.

For the saliency detector, we utilize MIDD [Tu *et al.*, 2021] directly for experimentation without any modifications. Our framework allows for the integration of existing models that are designed for dual-stream input tasks, such as RGB-D and RGB-T, including VST [Liu *et al.*, 2021] and CAVER [Pang *et al.*, 2023]. In fact, models originally developed for RGB

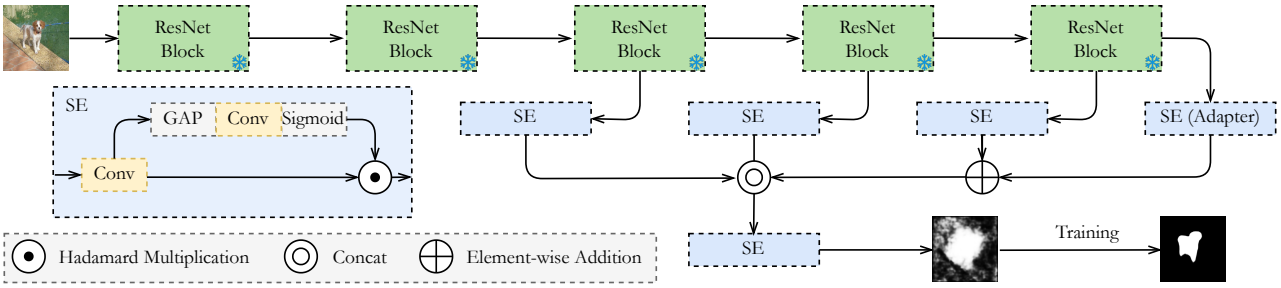

Figure 1: Architecture of the saliency cue extractor.

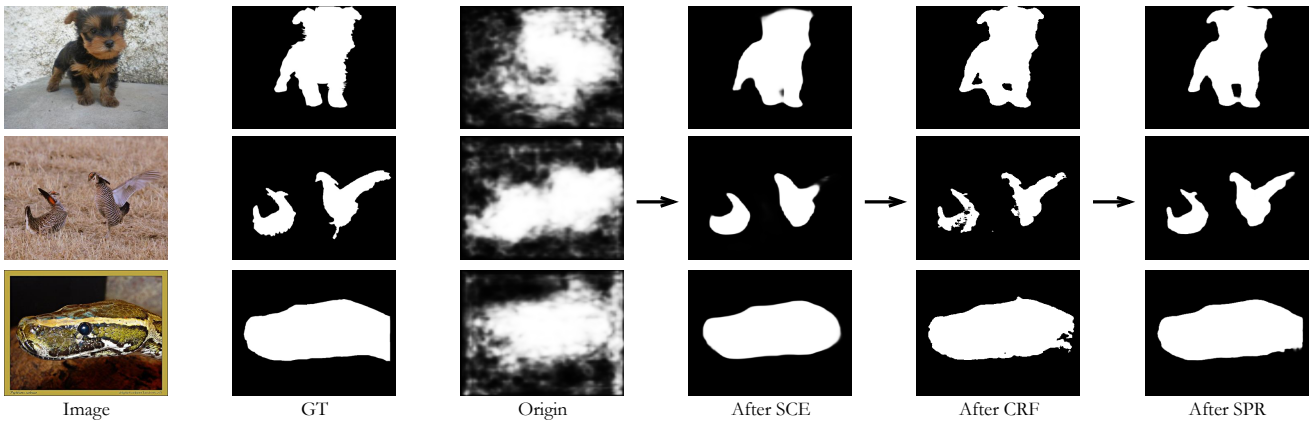

Figure 2: The optimization process of pseudo labels.

SOD tasks using only RGB input can also be employed. However, in such cases, we are unable to obtain effective saliency references from the modal input. It is worth mentioning that models that perform better under supervised training may not necessarily exhibit the same superiority under our unsupervised training. This discrepancy can be attributed to two factors: 1. We use pseudo labels instead of ground truth for training, and pseudo labels inherently contain some level of noise. Therefore, the model's ability to resist noise interference will affect its performance. 2. Our Self-rectify Pseudo-label Refinement (SPR) mechanism involves refining pseudo labels based on the model's outputs. If the model's outputs not only fail to suppress the noise in pseudo labels but also introduce new erroneous predictions, then the model's performance will deteriorate with further training.

3 Form high-quality pseudo-labels from scratch

In this section, we present the process of generating pseudo labels from scratch. As stated in the main text, the initial stage of our approach involves training a saliency cue extractor (SCE) to extract saliency cues from a pre-trained deep network. This procedure gradually transforms the output of the SCE from cluttered and unordered activation maps into saliency cues that accurately reflect the positions of salient objects, as illustrated in Figure 2.

To obtain activation maps that closely resemble those of the deep network, we must maintain a simple design for the SCE, resulting in relatively poor structural quality of the extracted saliency cues. Therefore, we employ Conditional Random Fields (CRF) to enhance these saliency cues and obtain initial pseudo labels. Subsequently, we train an additional saliency detector (SD) using these pseudo labels. CRF serves as a common post-processing technique that utilizes prior information about the image to compensate for the suboptimal structural quality of the saliency cues. Nevertheless, due to the absence of guidance from high-level semantic information and the influence of complex environments, the initial pseudo-labels obtained through this process often contain noise. Therefore, during the training of SD, we leverage the high-level semantic information present in the output of SD to rectify erroneous predictions within the pseudo labels.

Here, we would like to discuss further the prior and posterior rectifications incorporated in the proposed Self-rectify Pseudo-label Refinement. The posterior rectification pertains to the output of the saliency detector, while the prior rectification can refer to either CRF or the real-time pixel refiner adopted in our methodology. We apply CRF only once to acquire the initial pseudo labels, and all subsequent prior corrections are performed using the pixel refiner. This choice is justified by the time-consuming nature of CRF, as its processing duration for an image considerably exceeds the time required for the model to make predictions on the same image. Following the SPR process, the quality of the pseudo labels is further improved, as demonstrated in Figure 2.

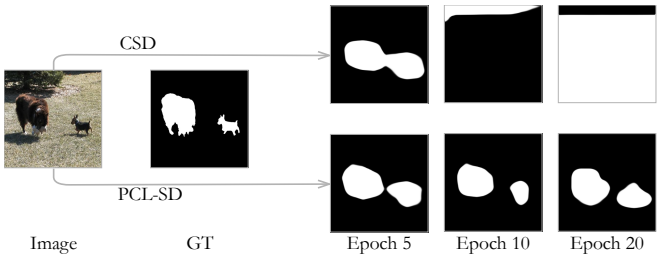

Figure 3: The comparison between proposed PCL-SD and CSD.

4 Curriculum Learning and SOD

To the best of our knowledge, we are the first to introduce the concept of curriculum learning into the field of salient object detection (SOD). Curriculum learning, originating from NLP tasks, centers around the idea of learning simple knowledge before tackling more challenging concepts. Designing a curriculum learning method involves focusing on three crucial aspects: (1) defining hard samples, (2) gradually incorporating hard samples, and (3) dealing with hard samples. In the main text, we discuss the first two points when presenting our proposed PCL-SD mechanism. For the third aspect, there are two strategies: a hard partition strategy, which strictly removes all hard samples by excluding them from backpropagation, and a soft partition strategy, which initially assigns a lower learning weight to the hard samples and gradually increases it. The latter still suffers from the problem of error accumulation caused by hard samples, leading to unstable training. Therefore, we adopt the former strategy, which rigidly removes all hard samples.

We conducted comparative experiments between our proposed PCL-SD and CSD [Zhou *et al.*, 2023b]. Surprisingly, we discovered that models trained with CSD are at risk of pattern collapse. As shown in Figure 3, at epoch 10, the model's outputs became meaningless and unrelated to the input images, consisting of large areas of pure white or black. We cannot solely attribute the phenomenon of pattern collapse to the interference of hard samples. However, when applying our proposed PCL-SD, the model training becomes more stable. Thus, we can conclude that PCL-SD effectively mitigates the problem of error accumulation caused by hard samples in the early stages of training, resulting in a more stable and robust training process.

In addition, we have also attempted to apply this hard sample handling strategy to supervised training. In this case, the determination of whether a specific point is a hard sample is based on its corresponding loss value, with a higher loss indicating increased difficulty. Regrettably, this approach did not yield any additional improvements. We suppose that two factors may contribute to this outcome: 1) the removal of hard samples impedes the model's ability to perceive intricate regions, and 2) the removal of hard samples introduces a disconnection between samples. To address the first factor, when designing PCL-SD, we only restrict the involvement of hard samples in gradient backpropagation during the initial few epochs of training, thereby avoiding the model's inability to fully learn from the knowledge contained within hard samples. As for the second factor, when calculating the loss using L_{sd}, each sample is treated as independent thereby no disconnection occurs. Overall, we believe that if these two factors can be avoided, curriculum learning still has good potential in SOD field.

5 More comparison results

We present visual comparisons of our proposed unsupervised method with other state-of-the-art methods on four tasks: RGB-D SOD (Figure 4), RGB-T SOD (Figure 5), video SOD (Figure 6), and Remote Sensing Image (RSI) SOD (Figure 7). For RGB-D SOD, the compared methods include VST [Liu *et al.*, 2021], CCFE [Liao *et al.*, 2022], DSU [Ji *et al.*, 2022], and TSD [Zhou *et al.*, 2023b]. For RGB-T SOD, the compared methods include MIDD [Tu *et al.*, 2021], CCFE, SRS [Liu *et al.*, 2023], and TSD. For video SOD, the compared methods include PCSA [Gu *et al.*, 2020], STVS [Chen *et al.*, 2021], WSVSOD [Zhao *et al.*, 2021], and TSD. For RSI SOD, the compared methods include LVNet [Li *et al.*, 2019] and MJRB [Tu *et al.*, 2022b]. The visual comparisons of RGB-D and RGB-T tasks demonstrate that our proposed method not only surpasses existing unsupervised methods but also achieves competitive performance compared to supervised methods. Moreover, for video SOD and RSI SOD tasks, we effectively leverage shared saliency knowledge transferred from the Nature Still Image (NSI) SOD tasks, resulting in impressive performance without any annotation data. These visual results confirm the outstanding performance of our proposed USOD framework and demonstrate the effectiveness of the proposed knowledge transfer approach.

6 Limitations & Future works

In this paper, we propose a unified unsupervised framework for salient object detection (SOD) based on knowledge transfer. While our method has achieved remarkable results, there are still areas for further improvement. We highlight three key directions for future research.

6.1 From Modality-Agnostic to Modality-Informed

During the training of our saliency detector on Nature Still Image (NSI) data, the model treats different modalities, such as depth, thermal, and optical flow, in a modality-agnostic manner. This approach does not specifically differentiate the input modality, treating it as generic reference information. While this enhances the model's generalization and robustness, it may limit the model's ability to exploit saliency information specific to each modality. Hence, transitioning from a modality-agnostic to a modality-informed approach is essential to facilitate better learning of shared saliency knowledge across different SOD tasks.

6.2 From Two-Stage to One-Stage

In this work, we adopt a two-stage framework, similar to previous unsupervised SOD (USOD) methods. This involves obtaining saliency clues or pseudo labels in the first stage and training a saliency detector using these labels in the second

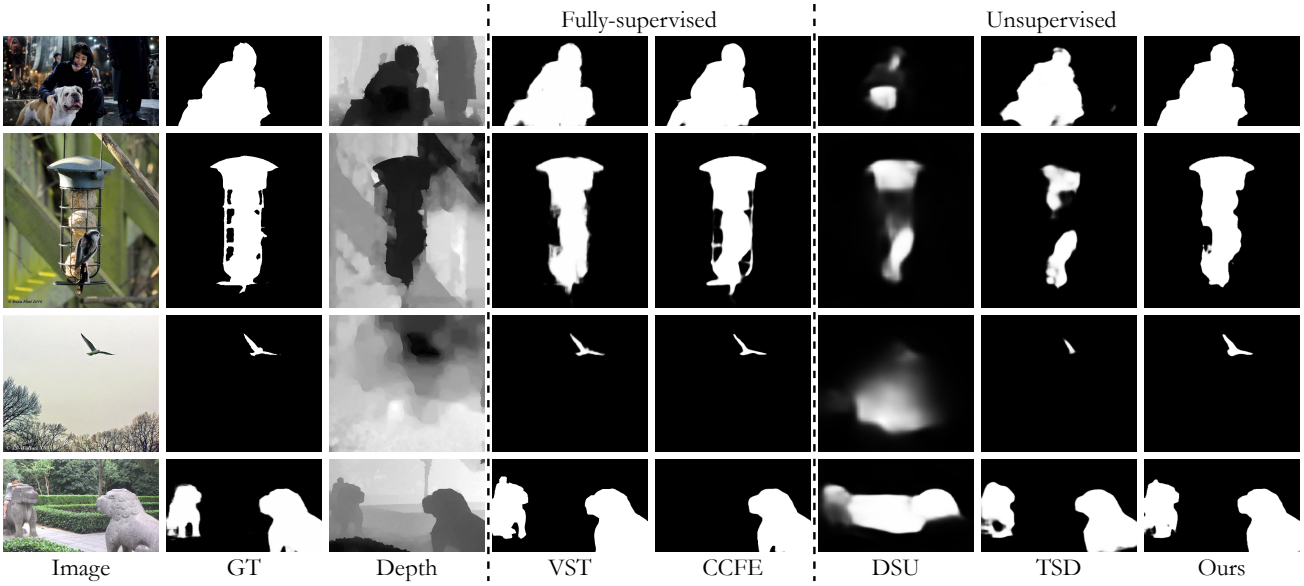

Figure 4: Visual comparison between the proposed method and the other state-of-the-art SOD methods on RGB-D SOD datasets.

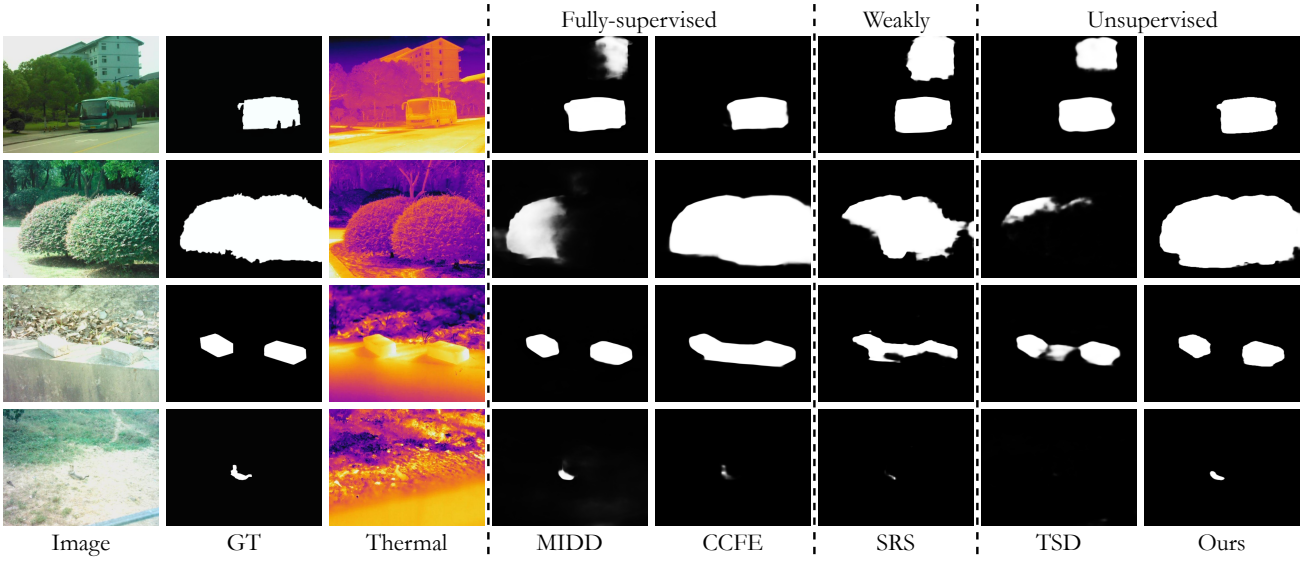

Figure 5: Visual comparison between the proposed method and the other state-of-the-art SOD methods on RGB-T SOD datasets.

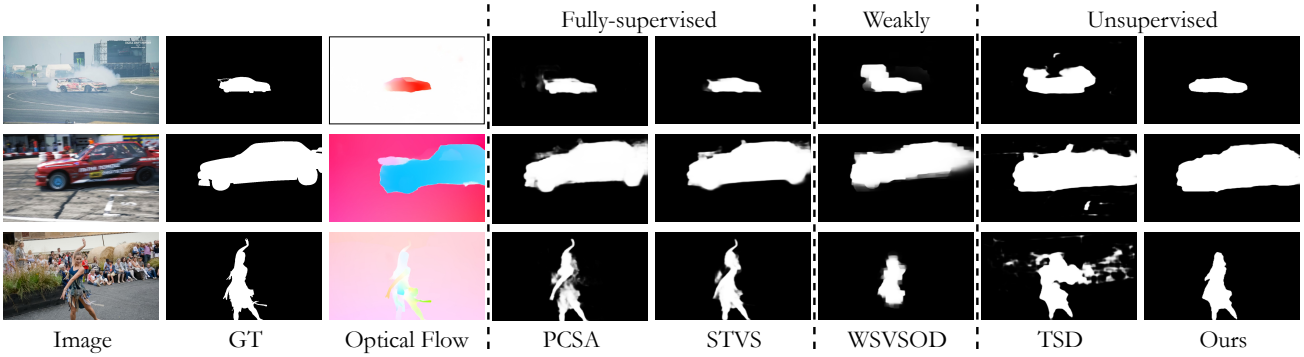

Figure 6: Visual comparison between the proposed method and the other state-of-the-art SOD methods on video SOD datasets.

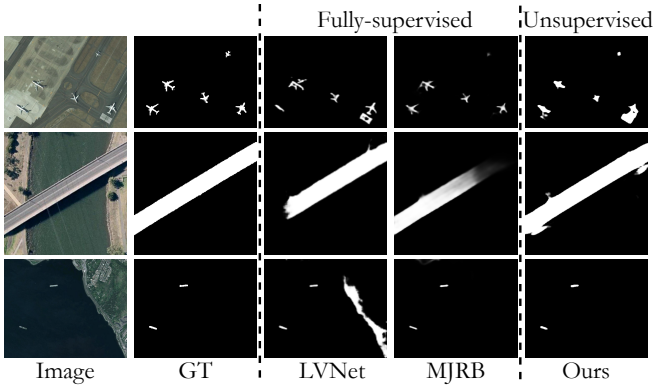

Figure 7: Visual comparison between the proposed method and the other state-of-the-art SOD methods on RSI SOD datasets.

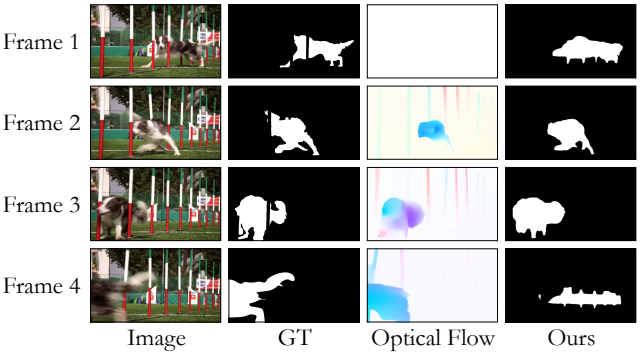

Figure 8: Our approach considers each frame within the video as an independent two-dimensional image for saliency detection. It effectively identifies salient objects in Frames 1, 2, and 3, yet encounters troubles in detecting salient objects in Frame 4.

stage. Although this two-stage training approach helps mitigate the instability associated with unsupervised methods, it may lead to potential disconnections. Mislocalized salient objects in the first stage are challenging to correct in the second stage. Therefore, exploring a one-stage model that addresses these issues is a worthwhile consideration.

6.3 Deeper and More Targeted Migration

In our work, we mainly focus on migrating to video SOD and RSI SOD. However, 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. Although this approach better utilizes saliency knowledge transferred from NSI SOD, it leads to issues depicted in Figure 8, where saliency is assessed based on the object's presence throughout the entire video, even if it is not salient in a specific frame. In future research, we plan to address these challenges and explore more tailored strategies specifically designed for these types of tasks.

References

[Chen *et al.*, 2021] Chenglizhao Chen, Guotao Wang, Chong Peng, Yuming Fang, Dingwen Zhang, and Hong Qin. Exploring rich and efficient spatial temporal interactions for real-time video salient object detection. *IEEE Transactions on Image Processing*, 30:3995–4007, 2021.

[Gu *et al.*, 2020] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. 2018.

[Ji *et al.*, 2022] Wei Ji, Jingjing Li, Qi Bi, Chuan Guo, Jie Liu, and Li Cheng. Promoting saliency from depth: Deep unsupervised rgb-d saliency detection. *arXiv preprint arXiv:2205.07179*, 2022.

[Li *et al.*, 2019] Chongyi Li, Runmin Cong, Junhui Hou, Sanyi Zhang, Yue Qian, and Sam Kwong. Nested network with two-stream pyramid for salient object detection in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):9156–9166, 2019.

[Liao *et al.*, 2022] Guibiao Liao, Wei Gao, Ge Li, Junle Wang, and Sam Kwong. Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7646–7661, 2022.

[Liu *et al.*, 2007] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[Liu *et al.*, 2021] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4722–4732, October 2021.

[Liu *et al.*, 2023] Zhengyi Liu, Xiaoshen Huang, Guanghui Zhang, Xianyong Fang, Linbo Wang, and Bin Tang. Scribble-supervised rgb-t salient object detection. *arXiv preprint arXiv:2303.09733*, 2023.

[Pang *et al.*, 2023] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Caver: Cross-modal view-mixed transformer for bi-modal salient object detection. *IEEE Transactions on Image Processing*, pages 1–1, 2023.

[Tu *et al.*, 2021] Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang. Multi-interactive dual-decoder for rgb-thermal salient object detection. *IEEE Transactions on Image Processing*, 30:5678–5691, 2021.

[Tu *et al.*, 2022a] Zhengzheng Tu, Yan Ma, Zhun Li, Chenglong Li, Jieming Xu, and Yongtao Liu. Rgbt salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia*, 2022.

[Tu *et al.*, 2022b] Zhengzheng Tu, Chao Wang, Chenglong Li, Minghao Fan, Haifeng Zhao, and Bin Luo. Orsi salient object detection via multiscale joint region and boundary model. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[Wang *et al.*, 2017] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[Wang *et al.*, 2023] Kunpeng Wang, Chenglong Li, Zhengzheng Tu, and Bin Luo. Unified-modal salient object detection via adaptive prompt learning, 2023.

[Zhao *et al.*, 2021] Wangbo Zhao, Jing Zhang, Long Li, Nick Barnes, Nian Liu, and Junwei Han. Weakly supervised video salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16826–16835, 2021.

[Zhou *et al.*, 2023a] Huajun Zhou, Peijia Chen, Lingxiao Yang, Xiaohua Xie, and Jianhuang Lai. Activation to saliency: Forming high-quality labels for unsupervised salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2):743–755, 2023.

[Zhou *et al.*, 2023b] Huajun Zhou, Bo Qiao, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. Texture-guided saliency distilling for unsupervised salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7257–7267, 2023.