Structured Cross-Resolution Distillation for Remote Sensing Salient Object Detection

Yanfeng Liu, *Student Member, IEEE,* Xin Zhang, Wei Guo, and Lefei Zhang, *Senior Member, IEEE*

***Abstract*—Existing salient object detection methods for optical remote sensing images achieve superior results with high-resolution inputs but exhibit significant degradation in low-resolution conditions. To bridge this resolution discrepancy, we propose a structured cross-resolution knowledge distillation (SCRKD) framework designed for severely low-resolution inputs. It leverages high-resolution models as teachers to guide low-resolution students through three synergistic distillation mechanisms: 1) multi-view correlation distillation, 2) multi-scale feature distillation, and 3) decoupled saliency distillation. In addition, we present Cascaded SCRKD that progressively refines structured knowledge in a multi-stage manner, achieving further performance boosts. Experiments on three datasets indicate that SCRKD surpasses 13 state-of-the-art methods across various cross-resolution settings. Besides, our framework based on three distinct baselines validates its model-agnostic nature. This work provides an efficient solution for low-resolution salient object detection. Code is available at *https://github.com/lyf0801/SCRKD*.**

***Index Terms*—Salient object detection, optical remote sensing image, cross-resolution distillation, multi-view correlation, decoupled saliency distillation.**

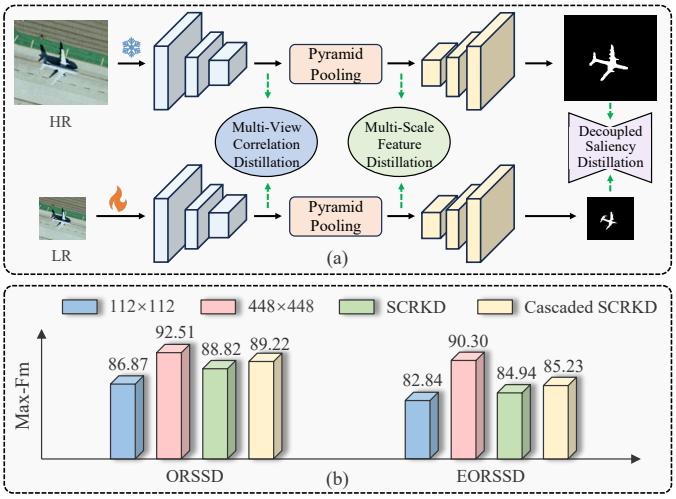


Fig. 1. (a) Illustration of the proposed SCRKD framework for RSI-SOD. (b) Max-Fm score comparison of on ORSSD and EORSSD datasets for 112×112 student, 448×448 teacher, SCRKD, and Cascaded SCRKD, respectively.

I. Introduction

SALIENT object detection in optical remote sensing images (RSI-SOD) aims to localizing geospatial regions of interest and man-made objects. As a fundamental preprocessing task in remote sensing [1], it facilitates downstream applications such as super-resolution [2], object detection [3], and scene understanding [4]. In recent years, research efforts in both computer vision and remote sensing communities have focused on optimizing RSI-SOD for real-time execution.

As for SOD in natural scenarios, early approaches provide several typical solutions to design lightweight SOD models and reduce the inference overhead [5], such as introducing depthwise separate convolution, and designing flexible self-adaptive operators, etc. Recently, Wang et al. [6] first present a backbone network specialized for SOD that eliminates ImageNet pre-training, which greatly alleviates feature complexity. However, these research efforts for natural scenes cannot directly perform well in remote sensing due to the distinct challenges of optical RSIs. To tackle the above problem, there are several research studies aiming at lightweight RSI-SOD [7], [8], [9]. All of the above-mentioned methods widely employ lightweight backbone networks (e.g., MobileNet, RepVGG) for multilevel feature extraction, reducing both model parameters and computational costs while accelerating inference. For example, Li et al. [9] introduce RepVGG-A0 with a small number of parameters to extract multiscale features, and utilize lightweight group attention and enhanced dynamic encoding module to boost spatial and channel attention information. Compared to ResNet-based saliency models, the above approaches definitely show an advantage in reasoning speed and maintain considerable detection performance. However, existing frameworks typically require high-resolution inputs for sufficient contextual knowledge, yet fail to tackle low-resolution RSIs. Particularly, Liu et al. [2] interestingly explore low-resolution RSI-SOD and propose a novel approach to distill cross-task knowledge from super-resolution to SOD. Nevertheless, the above work merely accounts for low-resolution settings of 224×224 and neglects to investigate extremely low-resolution scenarios such as 112×112 and 56×56.

As shown in Fig. 1(b), when deploying PSPNet [10] as the baseline network, its 448×448 version could achieve Max-

Manuscript received 14 March 2025; revised 15 June 2025, 16 July 2025 and 4 August 2025; accepted 11 August 2025. This work was supported by the National Natural Science Foundation of China under Grant 62431020, the Fundamental Research Funds for the Central Universities under Grant 2042025kf0030, and Seed Funding Project of Multisensor Intelligent Detection and Recognition Technologies R&D Center of CASC. *(Corresponding author: Lefei Zhang.)*

Yanfeng Liu and Lefei Zhang are with the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China (email: liuyanfeng99@whu.edu.cn; zhanglefei@whu.edu.cn).

Xin Zhang and Wei Guo are with the Multisensor Intelligent Detection and Recognition Technologies R&D Center, China Aerospace Science and Technology Corporation, Chengdu 610100, China (email: aley168@126.com; guowei3158@126.com).

Digital Object Identifier 10.1109/TGRS.2025.xxxxxxx

Fm scores of 92.51% and 90.30% on ORSSD and EORSSD, respectively. However, if the training scale is reduced to 112×112, the Max-Fm indicators of the same model decay from 92.51% to 86.87% and from 90.30% to 82.84% for both datasets. We attribute this degradation to the substantial domain gap between low-resolution and high-resolution RSIs, where low-resolution images inherently lack adequate spatial-contextual knowledge, causing severe deterioration in deep feature representation. Consequently, existing models struggle to accurately identify salient objects against complex background with irregular topology. This resolution deficiency creates significant challenges for existing methods in handling low-resolution scenes, which significantly hinders the deployment of neural networks for practical RSI analysis, posing urgent challenges to the remote sensing community.

To address the above research issues, we present a structured cross-resolution knowledge distillation (SCRKD) framework and an enhanced version, Cascaded SCRKD, to efficiently transfer abundant cross-resolution knowledge for RSI-SOD. Typically, three synergistic distillation strategies are proposed in Fig. 1(a), i.e., multi-view correlation distillation (MVCD), multi-scale feature distillation (MSFD), and decoupled saliency distillation (DSD). **Firstly**, MVCD defines three representations to exploit height-, width-, and channel-wise correlation information across resolutions in an omni-dimensional manner. **Secondly**, we compensate for cross-resolution discrepancies between student and teacher models via MSFD to adaptively optimize scale-specific low-resolution features. **Thirdly**, we propose DSD to reformulate saliency prediction as a dual-category segmentation task, and perform KL divergence over the channel dimensions to constrain the cross-resolution distribution variance. As shown in Fig. 1(b), SCRKD boosts Max-Fm of 112×112 PSPNet from 86.87% to 88.82% and 82.84% to 84.94% on both datasets. Additionally, its cascaded version further increases the student's performance. To our knowledge, this work is the first systematic study on cross-resolution distillation for RSI-SOD.

Furthermore, we evaluate the proposed SCRKD framework based on PSPNet [10], SegFormer [11], and TransXNet [12]. Experiments under diverse cross-resolution settings validate its model-agnostic nature. During inference, neither the teacher nor the assistant model is required, eliminating additional computational overhead. Thus, this study provides an alternative approach for efficient RSI-SOD instead of elaborate lightweight models, offering a promising research direction.

The main contributions of this work are listed as follows.

1) We investigate an efficient cross-resolution distillation framework, SCRKD, for RSI-SOD, offering valuable insights for remote sensing dense prediction.
2) We design three synergistic cross-resolution distillation strategies from omni-dimensional correlation, spatial pyramid, and saliency prediction levels.
3) To progressively leverage high-resolution knowledge, we construct a multi-stage Cascaded SCRKD framework that delivers further performance gains.

The remaining article is organized as follows. Section II presents related studies of RSI-SOD and knowledge distillation. We describe the methodology of the proposed SCRKD in Section III and discuss experiments in Section IV. Finally, Section V draws a conclusion.

II. Related Work

A. Remote Sensing Salient Object Detection

RSI-SOD has garnered significant attention due to its critical role in various remote sensing applications. Over the past decade, numerous algorithms have been proposed to address the unique challenges of RSI-SOD, including complex backgrounds, varying object scales, diverse spatial resolutions, and adversarial defenses. These methods can be broadly categorized into traditional approaches and deep learning-based methods. In this subsection, we provide the brief review of deep RSI-SOD models in recent years.

Initially, a number of research efforts have developed some large-scale public datasets for RSI-SOD [13], [14], [15], [16], and then numerous deep models have been proposed in rapid succession. Early approaches mostly concern with multi-scale feature fusion, edge guidance, local and global collaborative learning, multiple attention mechanisms, lightweight models, and so on. For instance, Liu et al. [17] introduce an efficient global context strategy to compensate for the local spatial features of convolutional networks. Xie et al. [18] employ boundary features to guide channel attention to salient edges and maintain spatial details. In addition, Wang et al. [19] present an mutually supervised bootstrap loss for edge and saliency to enhance the learning of irregular topologies and complex edges. Recent research studies have focused on typical novel issues in RSI-SOD, such as adversarial attacks, uncertainty, ensemble learning, novel network structures, and pseudo-label contrastive learning. For example, Yan et al. [20] propose to combine convolutional operators and self-attention mechanisms, and design a heterogeneous adaptive semantic model for this topic. Liu et al. [21] propose an integrated and detailed ensemble learning framework, which addresses the imbalance between deep and shallow features and effectively preserves object integrity and edge details.

While existing studies have made substantial contributions to this field, their reliance on high-resolution inputs renders them ineffective under low-resolution conditions. In this article, we investigate cross-resolution distillation for low-resolution RSI-SOD, paving the way for solving the aforementioned challenges.

B. Knowledge Distillation in Remote Sensing

In recent years, the remote sensing community has proposed knowledge distillation-based studies for various downstream tasks, such as object detection [22], semantic segmentation [23], change detection [24], scene understanding [25], and land cover classification [26]. For instance, Li et al. [27] combine feature-based, relation-based, and instance-aware distillation methods for efficient remote sensing object detection. Dong et al. [23] introduce a distilling segmenter framework for semantic segmentation of RSIs, leveraging channel-weighted attention-guided feature distillation and target-nontarget distillation strategies. Typically, Pang et al. [24] introduce a hierarchical correlation distillation framework for change detection

across image pairs of varying quality, transferring knowledge from high-quality to low-quality samples. In addition, Zhang et al. [3] present the super-resolution generative distillation and cross-modality affinity distillation to leverage knowledge from RGB modality for thermal small-object detection.

To summarize, these different approaches are all based on knowledge distillation, which leverage multilevel knowledge from complicated teacher models to achieve compact student models for remote sensing image processing.

C. Cross-Resolution Knowledge Distillation

Existing literature on cross-resolution distillation [28] primarily focuses on low-resolution face recognition [29], low-resolution object detection [30], video processing [31], etc.

Typically, Shin et al. [32] propose an attention similarity knowledge distillation approach to enhance low-resolution face recognition by transferring attention maps from a high-resolution teacher network. Zhu et al. [33] present scale-aware knowledge distillation framework, a novel approach to improve small object detection through a scale-decoupled feature distillation module and a cross-scale assistant. Ma et al. [31] improve video recognition accuracy on low-resolution frames by addressing the mismatch between network architecture and input scale. Furthermore, Guo et al. [34] extend distillation to the input level, enabling flexible cost control by adjusting both network architecture and image quality, and introduce an input spatial representation distillation mechanism for image classification and object detection tasks. Recently, Wang et al. [35] propose a multi-scale cross distillation method, which combines multi-scale training to enable single-scale inference for aerial object detection, and integrates adaptively cross-scale knowledge through a parallel multi-branch architecture.

However, existing methods suffer from three critical limitations: 1) hard to jointly explore structured cross-resolution knowledge at correlation-, spatial-, and logit-level; 2) ineffective knowledge transfer on imbalanced saliency prediction distributions; 3) lack of exploration for ultra-low-resolution student scenarios (e.g., 112×112 or 56×56 inputs). To this end, we firstly introduce cross-resolution distillation into RSI-SOD and propose an efficient distillation framework tailored for low-resolution RSIs, compensating for the drastic performance degradation caused by insufficient spatial resolution.

III. Methodology

This section describes the proposed SCRKD in detail. The core techniques of SCRKD are discussed elaborately.

A. Overview of SCRKD

As illustrated in Fig. 2, the proposed SCRKD consists of a well-trained teacher model, a learnable student model, and three designed synergistic distillation modules: MVCD, MSFD, and DSD. Specifically, the teacher model takes high-resolution RSIs as input and generates high-resolution saliency maps, while the student model feeds low-resolution RSIs and receives multi-level structured knowledge from the high-resolution teacher model to predict refined saliency maps. During the training phase, the student model incorporates distillation losses from MVCD, MSFD, and DSD, along with the primary saliency detection loss, which is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{SOD} + \alpha \cdot \mathcal{L}_{MVCD} + \beta \cdot \mathcal{L}_{MSFD} + \gamma \cdot \mathcal{L}_{DSD}, \quad (1)$$

where $\mathcal{L}_{MVCD}$, $\mathcal{L}_{MSFD}$, and $\mathcal{L}_{DSD}$ denote losses of MVCD, MSFD, and DSD, respectively. $\mathcal{L}_{SOD}$ represents saliency loss, which combines a binary cross-entropy (BCE) loss and a weighted intersection over union (wIoU) loss as follows:

$$\mathcal{L}_{SOD} = \sum_{i=1}^{n} (L_{bce}(S_p^i, S_g) + L_{wiou}(S_p^i, S_g))/2^{i-1}, \quad (2)$$

where S_p and S_g indicate predicted and ground-truth saliency maps, and n is the number of detection heads for supervision.

B. Feature Disentanglement and Adaptive Aggregation

The proposed MVCD, MSFD, and DSD facilitate the transfer of complementary structured knowledge from the self-correlation level, multiscale feature level, and saliency prediction level, respectively. Among them, both MVCD and MSFD distill knowledge to the multi-scale features of the student model. However, the standalone student feature struggles to consistently and cohesively capture low-resolution scale-specific knowledge, scale-invariant correlation representations, and high-resolution fine-grained features simultaneously. To address this limitation, we introduce a simple yet effective feature disentanglement and adaptive aggregation mechanism.

As shown in Fig. 2, for the feature of student at a given stage, we employ three parallel convolutional layers to perform adaptive transformation, yielding three distinct feature embeddings. These representations are used for: (1) the student's independent low-resolution saliency feature representation, (2) learning the correlation distillation knowledge from the teacher model, and (3) learning the spatial feature knowledge from the teacher model. This process can be expressed as:

$$f_S', \ f_{MVCD}, \ f_{MSFD} = \mathcal{F}_{Dis}(f_S), \quad (3)$$

where $\mathcal{F}_{Dis}(\cdot)$ denotes the function of feature disentanglement procedure, and f_S indicates the scale-specific low-resolution student feature. f_S', f_{MVCD}, and f_{MSFD} represent the above-mentioned three distinct feature embeddings.

After the distillation training of SCRKD, we introduce an convolution-based attention method, as shown in "Adaptive Aggregation" in Fig. 2, to combine the student's preserved low-resolution feature, the distilled correlation feature, and the distilled high-resolution spatial feature, as follows:

$$w_1, w_2, w_3 = \sigma(\mathcal{C}_{3\times3}([f_{LR}', f_{MVCD}, f_{MSFD}])), \quad (4)$$

where $\mathcal{C}_{3\times3}(\cdot)$ indicates the standard 3×3 convolution for attention weights generation, $[\cdot,\cdot]$ denotes channel-wise concatenation, and $\sigma(\cdot)$ represents the Sigmoid activation function.

Then, we could utilize these dynamic weights to aggregate the three disentangled embeddings as a unified vector, i.e.,

$$f_S'' = \mathcal{C}_{3\times3}(w_1 \cdot f_S' \oplus w_2 \cdot f_{MVCD} \oplus w_3 \cdot f_{MSFD}), \quad (5)$$

where $\oplus$ denotes the element-wise summation, and f_S'' is the embedding for student saliency head to yield final predictions.

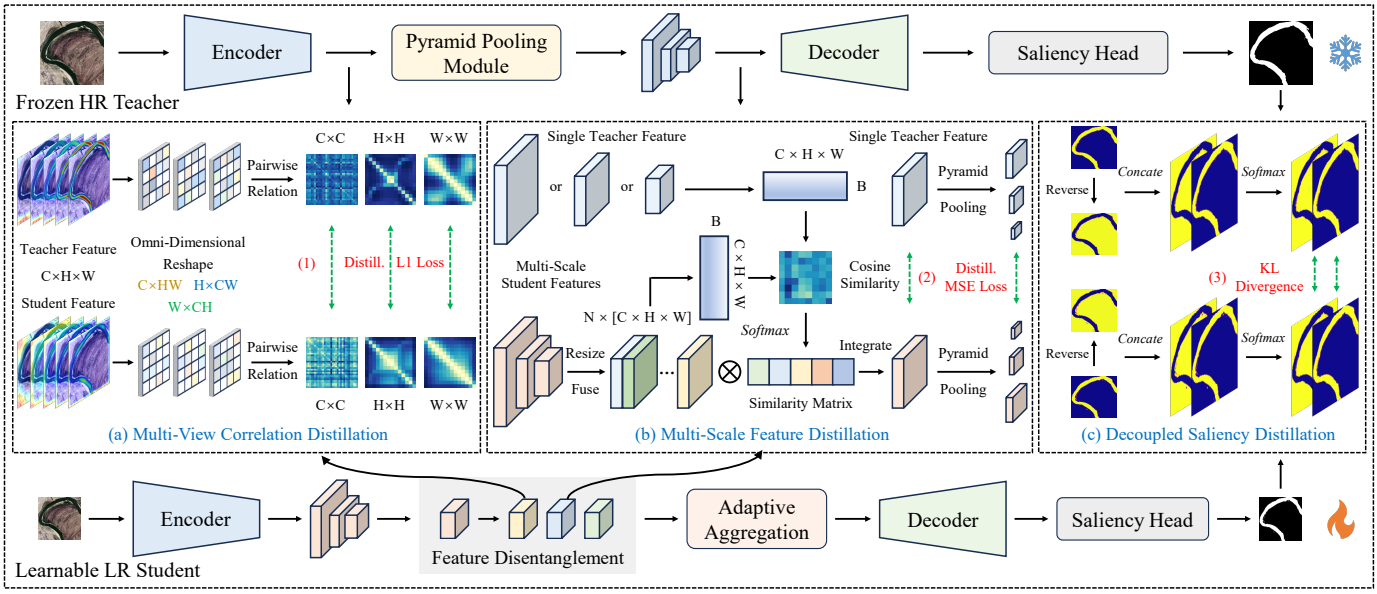


Fig. 2. Illustration of the proposed SCRKD, our three novel distillation modules are presented in detail: (a) multi-view correlation distillation (MVCD); (b) multi-scale feature distillation (MSFD); (c) decoupled saliency distillation (DSD). Pyramid pooling module (PPM) is presented in our baseline PSPNet [10]. Feature disentanglement and adaptive aggregation are discussed in Section III-B). Note that we eliminate PPM for student illustration for a better view.

C. Multi-View Correlation Distillation (MVCD)

Global-level scene-invariant representation is a crucial property for characterizing diverse RSIs with various salient objects [4], as they remain consistent regardless of variations in spatial scales. Existing distillation methods predominantly emphasize channel-wise correlations [36] or cross-sample relationships [37], always neglecting the inherent multi-level correlations within the features. Moreover, the efficient measurement of multi-dimensional global structured correlation for the RSI-SOD task remains an unresolved challenge. To address these limitations, we introduce a cross-resolution multi-view correlation distillation approach, referred to as MVCD.

As illustrated in Fig. 2(a), we propose an omni-dimensional feature transformation method to characterize saliency knowledge across different dimensions, namely the channel level, height level, and width level. These transposed vectors from different dimensions capture complementary global factors, playing a crucial role in representing salient objects. For a teacher-student feature pair $f_T, f_{MVCD} \in \mathbb{R}^{C\times H\times W}$, the comprehensive transformation method generates three reshaped features: $S_c \in \mathbb{R}^{C\times HW}$, $S_h \in \mathbb{R}^{H\times CW}$, $S_w \in \mathbb{R}^{W\times CH}$ for the student and $T_c \in \mathbb{R}^{C\times HW}$, $T_h \in \mathbb{R}^{H\times CW}$, $T_w \in \mathbb{R}^{W\times CH}$ for the teacher, denoting the reshaped features along the channel, height, and width dimensions, respectively.

Based on the aforementioned transposed features, we can employ cosine similarity to compute omni-dimensional self-correlation matrices for both the teacher and student spatial features. For instance, for the student, its channel-wise correlation matrix can be expressed as:

$$C_S = \frac{S_c \cdot S_c^{\mathrm{T}}}{\|S_c\|_2 \cdot \|S_c^{\mathrm{T}}\|_2} \in \mathbb{R}^{C\times C}, \tag{6}$$

where $\|\cdot\|_2$ indicates the Euclidean normalization, and C_S represents the channel-wise correlation map of student. Similarly, we can obtain height-wise and width-wise correlation matrices H_S and W_S for the student, as well as the three-dimensional self-correlation matrices C_T, H_T, and W_T for the teacher.

To enable the low-resolution student to learn structured knowledge from the high-resolution teacher effectively, we employ a straightforward yet efficient optimization function. This function guides the student to mimic the scale-invariant, omni-dimensional correlation knowledge, formulated as:

$$L_{MVCD} = ||C_S - C_T||_1 + ||H_S - H_T||_1 + ||W_S - W_T||_1, \tag{7}$$

where $||\cdot||_1$ denotes L1 loss function. Through this comprehensive global correlation guidance, the student is able to capture scale-invariant abstract semantic information from the teacher, thereby exploiting more valuable clues for RSI-SOD.

D. Multi-Scale Feature Distillation (MSFD)

High-resolution RSIs can extract more discriminative activations for salient objects, offering valuable hints for low-resolution students to learn fine-grained features. However, the huge resolution gap between high-resolution teachers and low-resolution students makes directly one-to-one stage-wise feature matching impractical [38]. While aligning features with identical spatial scales is possible, this approach inevitably omits some critical high-resolution contextual knowledge, limiting the capacity of feature distillation [39]. To address this issue, we propose to integrate multi-scale low-resolution student features to adaptively incorporate high-resolution contextual guidance from the teacher. This ensures all stages of student and teacher features participate in distillation, maximizing the potential of cross-resolution spatial feature transfer.

As presented in Fig. 2(b), to distill knowledge from the high-resolution teacher feature f_T at a specific stage, MSFD first employs bilinear upsampling, adaptive average pooling,

or an identity mapping to align the resolutions of the multi-scale low-resolution student features with f_T. Subsequently, a convolution is utilized to refine the aligned features f_{com}, i.e.,

$$f_{com} = \mathcal{C}_{3\times3}([\mathcal{F}_{align}(f^1_{MSFD}),\ ...,\ \mathcal{F}_{align}(f^n_{MSFD})]), \quad (8)$$

where $\mathcal{F}_{align}(\cdot)$ denotes the spatial alignment function mentioned above, and n indicates the number of multi-scale stages (e.g., $n = 5$ for PSPNet [10] and $n = 4$ for SegFormer [11]).

Then, we perform the dot product function between the combined student embeddings f_{com} and the high-resolution teacher feature f_T to generate the affinity map $\mathcal{A}$ as follows:

$$\mathcal{A} = \frac{f_{com} \cdot f_T^{\mathrm{T}}}{\tau_1}, \quad (9)$$

where τ_1 is a learnable parameter to control the sharpness of softmax logits during the training phase. Moreover, the sum-to-one strategy is introduced via a softmax function to calculate the integration weights for $f^1_{MSFD}, ..., f^n_{MSFD}$, i.e.,

$$\lambda_i = \frac{\exp(\mathcal{A}_i)}{\sum_{j=1}^{n}\exp(\mathcal{A}_j)}. \quad (10)$$

Therefore, a dynamically integrated student feature f_D can be obtained through an adaptive weighted summation, as follows:

$$f_D = \sum_{i=1}^{n} \lambda_i \otimes \mathcal{F}_{align}(f^i_{MSFD}), \quad (11)$$

where $\otimes$ indicates element-wise multiplication function.

As shown in Fig. 2(b), unlike traditional feature distillation losses that solely focus on absolute errors between features, we additionally introduce multi-scale global pyramid pooling to guide the student in learning abstract patterns of the teacher feature. This approach aims to preserve scale-invariant contextual structured knowledge across resolutions. Consequently, the overall loss function of MSFD can be formulated as:

$$L_{MSFD} = ||f_T - f_D||_2 + \sum_{i=0}^{4} ||\mathcal{P}_{2^i}(f_T) - \mathcal{P}_{2^i}(f_D)||_2, \quad (12)$$

where $\mathcal{P}_{2^i}(\cdot)$ is the global average pooling operation with an output size of $2^i \times 2^i$, and $||\cdot,\cdot||_2$ denotes the L2 loss function.

E. Decoupled Saliency Distillation (DSD)

The original logit distillation utilizes a temperature-scaled softmax normalization to produce a unified logit probability distribution, followed by KL divergence for distillation. Nowadays, this approach has been generalized to other tasks, e.g., image segmentation [40] and multi-category classification [41]. However, these tasks distinct significantly from SOD, as they involve predictions for multiple semantic categories, whereas SOD generates a single-category saliency map. As a result, traditional logits distillation can only be applied along the spatial dimensions of $H \times W$. However, applying softmax to a single-category saliency map with a large spatial size, where predictions vary significantly across pixels, results in the loss of discriminative information, such as extremely high and low activation values, making it impractical for RSI-SOD. To address this deficiency, we propose to leverage non-salient, non-target dark knowledge and design a decoupled saliency logits distillation method, termed DSD.

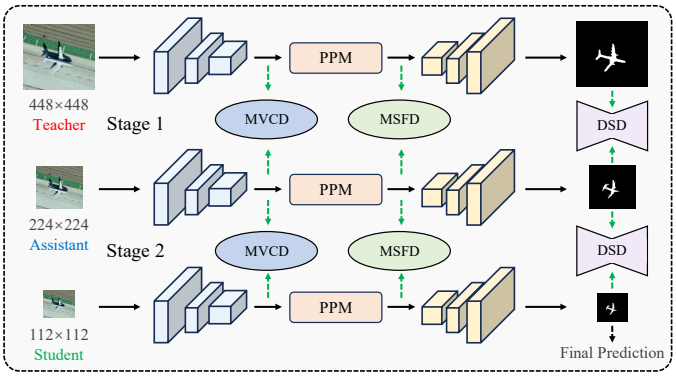


Fig. 3. Illustration of Cascaded SCRKD, which distills high-resolution teacher knowledge to intermediate-resolution assistant in the first stage, and then distills assistant knowledge to low-resolution student in the second stage.

As illustrated in Fig. 2(c), DSD first defines a non-saliency representation, which we believe also contains hidden knowledge beneficial for RSI-SOD. Specifically, for the saliency maps S_S and S_T predicted by the student and teacher, we first apply the sigmoid function to eliminate negative activations and then subtract them from the all-ones matrix $\mathbf{J}$ to derive the non-saliency maps $\mathbf{J} - \sigma(S^S)$ and $\mathbf{J} - \sigma(S^T)$. Then, we utilize channel-wise concatenation to combine the saliency and non-saliency maps to obtain decoupled logits as follows:

$$D^S = [\sigma(S^S), \mathbf{J} - \sigma(S^S)],\ D^T = [\sigma(S^T), \mathbf{J} - \sigma(S^T)], \quad (13)$$

where D_S and D_T are two-category semantic normalized representations. Finally, we employ **KL** divergence for decoupled saliency and non-saliency distillation collaboratively, i.e.,

$$L_{DSD} = \frac{1}{W \cdot H}\sum_{i=1}^{W}\sum_{j=1}^{H} \mathbf{KL}(\psi(\frac{D^S_{i,j}}{\tau_2})||\psi(\frac{D^T_{i,j}}{\tau_2})), \quad (14)$$

where τ_2 is a temperature hyperparameter for soft logits, and $\psi(\cdot)$ denotes softmax function. W and H are the width and height for predicted saliency maps. Benefiting from the above design, DSD uniformly optimizes the cross-resolution saliency and non-saliency distribution variance in the output level.

F. Cascaded SCRKD

If there exists an extreme resolution gap between the teacher and student, such as 448×448 versus 112×112, directly applying the proposed SCRKD framework cannot fully unleash the potential of cross-resolution knowledge distillation. To this end, we introduce an assistant model with an intermediate resolution and propose a cascaded SCRKD framework in a two-stage distillation manner, as illustrated in Fig. 3.

Specifically, in the first distillation stage, we employ the high-resolution well-trained model as the teacher and the intermediate-resolution assistant model as the student, executing the SCRKD framework for distilling. After obtaining the well-trained assistant model in the first stage, we proceed to the second distillation stage, where the intermediate-resolution assistant model serves as the teacher and the low-resolution

model as the student, performing SCRKD distillation training once again. Through this two-stage progressive learning process, the low-resolution student model can capture more fine-grained cross-resolution knowledge, resulting in a student model with stronger generalization capabilities.

IV. Experiments

In this section, we introduce the experimental protocol, discuss the comparison analysis and ablation study, and present the typical visualization to reveal the effects of SCRKD.

A. Experimental Protocol

Here, we briefly introduce the dataset information, evaluation metrics, implementation details, and comparison methods.

1) Datasets: We conduct experiments on three widely-used RSI-SOD datasets to evaluate the cross-resolution distillation performance, i.e., ORSSD [14], EORSSD [15], and ORSI-4199 [16]. Specifically, the ORSSD dataset comprises 800 RSIs, with 600 allocated for training and 200 for testing. The EORSSD dataset consists of 2000 RSIs, of which 1400 are used for training and 600 for testing. Additionally, ORSI-4199 is a large-scale dataset containing 4199 images, divided into 2000 for the training subset and 2199 for the test subset.

2) Evaluation Metrics: Three metrics, namely mean absolute error (MAE), F-measure, and S-measure, are employed for the quantitative comparison. MAE evaluates the absolute pixel-level difference between the predicted saliency map and the ground-truth saliency map, i.e.,

$$\text{MAE} = \frac{1}{W \cdot H} \sum_{i=1}^{W} \sum_{j=1}^{H} |S_p(i,j) - S_g(i,j)|, \tag{15}$$

where S_p denotes the predicted saliency map, S_g represents the ground-truth saliency map, i and j are the pixel coordinates, and $W \cdot H$ is the total number of pixels for an RSI.

F-measure is utilized to balance the precision and recall in saliency detection, which is defined as:

$$F_\beta = \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}, \tag{16}$$

where β^2 is set to 0.3 as per the original configuration.

S-measure evaluates the structural similarity between the label and the prediction at both region and object levels:

$$S_m = \alpha \times S_o(S_p, S_g) + (1-\alpha) \times S_r(S_p, S_g), \tag{17}$$

where S_o denotes the object-aware similarity, S_r represents the region-aware similarity, and α is equal to 0.5.

3) Implementation Details: The experimental environment is Ubuntu 20.04 system and PyTorch 2.1 toolbox, and all experiments are deployed on a single NVIDIA GeForce RTX 3090 or 4090 GPUs. PSPNet [10] incorporates a five-stage encoder with ResNet50 pretraining weights. SegFormer [11] adopts its b1 version, with "patch-embed1" set to 2, and it includes a four-stage encoder with official pre-trained weights. TransXNet [12] employs its tiny variant, configuring an input stride of 2 and adopting a four-stage encoder architecture initialized with official pre-trained weights. Following previous work [2], [4], [19], [17], we conduct training and testing on three datasets separately, and employ data augmentation techniques are applied to the training sets. All saliency map metrics are computed at a resolution of 448×448. For all comparative methods and the proposed SCRKD, the AdamW optimizer is employed for training, with a total of 100 epochs, a learning rate of 2e-5, and a batch size of 8. To ensure balanced optimization, we set the loss coefficients to $\alpha = 1$, $\beta = 1$, and $\gamma = 10$ for PSPNet, $\alpha = 0.1$, $\beta = 1$, and $\gamma = 100$ for SegFormer, and $\alpha = 0.2$, $\beta = 0.6$, and $\gamma = 20$ for TransXNet, effectively aligning the magnitudes of the SOD and three distillation losses during training. The logit temperature is simply set to 1. During inference, both the teacher and the potential assistant in SCRKD can be removed, thus introducing no additional parameters.

4) Comparison Methods: To provide a comprehensive validation, we introduce 13 state-of-the-art algorithms for comparison, including six methods for general knowledge distillation: FitNet [38], AT [42], ReviewKD [43], SRD [44], LogitStdKD [45], and FAMKD [46], three approaches for dense prediction: CWD [36], SKD [40], and STONet [47], two algorithms for cross-resolution distillation: PD [34] and CMHRD [3], and two lightweight models: ISAANet [48] and SOLNet [9].

B. Preliminary Experiments

We establish three distinct baseline models for validation: 1) a CNN-based PSPNet [10], 2) a Transformer-based SegFormer [11], 3) a hybrid CNN-Transformer model, TransXNet [12]. To systematically evaluate RSI-SOD performance across varying resolution conditions, we conduct multi-scale experiments at 112×112, 224×224, and 448×448, with quantitative results presented in Table I. As can be observed, the quantitative results demonstrate consistent performance degradation on both ORSSD and EORSSD datasets as resolution decreases, confirming that lower resolutions significantly impair RSI-SOD performance. For the ORSI-4199 dataset, metrics at 448×448 and 224×224 exhibit mixed trends, which we attribute to the predominance of large-scale salient objects in this dataset. Consequently, the performance variance across resolutions is less pronounced compared to other datasets. Based on these substantial performance discrepancies, we define three cross-resolution distillation protocols: 1) from 224×224 to 112×112 for all three datasets; 2) from 448×448 to 224×224 for ORSSD and EORSSD; and 3) from 448×448 to 112×112 for ORSSD and EORSSD. To further investigate ultra-low-resolution distillation scenarios, we conduct experiments on the EORSSD dataset under three cross-resolution settings: distilling from 1) a 112×112 teacher to a 56×56 student, 2) a 224×224 teacher to a 56×56 student, and 3) a 448×448 teacher to a 56×56 student. Notably, SCRKD achieves a remarkable nearly 6% boost in F_β for the 56×56 student model, demonstrating its superior capability in extreme low-resolution distillation. Detailed quantitative comparisons for 56×56 students are provided on our GitHub website.

C. Comparison with the State-of-the-Art Algorithms

In this subsection, we firstly employ PSPNet [10] as the baseline for distillation comparison, then extend to SegFormer [11] and TransXNet [12] to validate the generalization ability.

TABLE I
QUANTITATIVE PERFORMANCE OF PSPNET AND SEGFORMER ON THREE RSI-SOD DATASETS WITH VARIOUS TRAINING RESOLUTIONS.

Networks	Publication	Input Size	ORSSD Dataset [14]			EORSSD Dataset [15]			ORSI-4199 Dataset [16]		
			MAE ↓	F_β ↑	S_m ↑	MAE ↓	F_β ↑	S_m ↑	MAE ↓	F_β ↑	S_m ↑
PSPNet [10]	CVPR 2017	112×112	0.0149	0.8687	0.8944	0.0098	0.8284	0.8796	0.0367	0.8368	0.8561
PSPNet [10]	CVPR 2017	224×224	0.0119	0.8981	0.9146	0.0077	0.8796	0.9135	**0.0344**	**0.8538**	0.8676
PSPNet [10]	CVPR 2017	448×448	**0.0104**	**0.9251**	**0.9309**	**0.0071**	**0.9030**	**0.9301**	0.0350	0.8465	**0.8689**
SegFormer [11]	NeurIPS 2021	112×112	0.0163	0.8681	0.8995	0.0090	0.8459	0.8947	0.0336	0.8499	0.8690
SegFormer [11]	NeurIPS 2021	224×224	0.0113	0.9139	0.9273	0.0070	0.8917	0.9235	**0.0307**	**0.8635**	0.8765
SegFormer [11]	NeurIPS 2021	448×448	**0.0106**	**0.9218**	**0.9314**	**0.0067**	**0.9108**	**0.9348**	0.0318	0.8631	**0.8786**
TransXNet [12]	TNNLS 2025	112×112	0.0133	0.8799	0.9099	0.0086	0.8471	0.8979	0.0336	0.8492	0.8679
TransXNet [12]	TNNLS 2025	224×224	0.0102	0.9238	0.9381	0.0064	0.9001	0.9296	0.0293	**0.8716**	0.8839
TransXNet [12]	TNNLS 2025	448×448	**0.0086**	**0.9297**	**0.9413**	**0.0053**	**0.9155**	**0.9408**	**0.0287**	0.8703	**0.8852**

TABLE II
QUANTITATIVE CROSS-RESOLUTION DISTILLATION PERFORMANCE ON THREE RSI-SOD DATASETS WITH 13 KNOWLEDGE DISTILLATION APPROACHES. THE TOP THREE RESULTS ARE MARKED IN RED, GREEN AND BLUE, RESPECTIVELY.

PSPNet [10]	Publication	Input Size	ORSSD Dataset [14]			EORSSD Dataset [15]			ORSI-4199 Dataset [16]		
			MAE ↓	F_β ↑	S_m ↑	MAE ↓	F_β ↑	S_m ↑	MAE ↓	F_β ↑	S_m ↑
Teacher	CVPR 2017	224×224	0.0119	0.8981	0.9146	0.0077	0.8796	0.9135	0.0344	0.8538	0.8676
Student	CVPR 2017	112×112	0.0149	0.8687	0.8944	0.0098	0.8284	0.8796	0.0367	0.8368	0.8561
+ FitNet [38]	ICLR 2016	112×112	0.0146	0.8787	0.9010	0.0090	0.8398	0.8890	0.0346	0.8403	0.8623
+ AT [42]	CVPR 2017	112×112	0.0129	0.8851	0.9096	0.0091	0.8428	0.8908	0.0345	0.8424	0.8639
+ CWD [36]	ICCV 2021	112×112	0.0138	0.8788	0.9039	0.0091	0.8395	0.8870	0.0356	0.8399	0.8597
+ ReviewKD [43]	CVPR 2021	112×112	0.0132	0.8860	0.9070	0.0089	0.8429	0.8929	0.0346	0.8447	0.8638
+ SKD [40]	TPAMI 2023	112×112	0.0140	0.8760	0.8987	0.0094	0.8391	0.8865	0.0354	0.8389	0.8569
+ SRD [44]	AAAI 2024	112×112	0.0150	0.8764	0.8995	0.0095	0.8388	0.8864	0.0358	0.8376	0.8588
+ LogitStdKD [45]	CVPR 2024	112×112	0.0140	0.8732	0.8999	0.0100	0.8329	0.8815	0.0354	0.8422	0.8626
+ FAMKD [46]	WACV 2024	112×112	0.0141	0.8745	0.8996	0.0089	0.8415	0.8898	0.0346	0.8436	0.8629
+ STONet [47]	TGRS 2024	112×112	0.0131	0.8802	0.9053	0.0095	0.8414	0.8878	0.0332	0.8421	0.8646
+ PD [34]	TPAMI 2024	112×112	0.0130	0.8817	0.9045	0.0097	0.8367	0.8844	0.0348	0.8407	0.8606
+ CMHRD [3]	TGRS 2024	112×112	0.0141	0.8844	0.9040	0.0092	0.8368	0.8859	0.0354	0.8405	0.8607
ISAANet [48]	TGRS 2024	112×112	0.0139	0.8804	0.9108	0.0101	0.8283	0.8861	0.0400	0.8191	0.8495
SOLNet [9]	TGRS 2025	112×112	0.0232	0.8307	0.8718	0.0139	0.8146	0.8710	0.0486	0.8030	0.8260
+ SCRKD (Ours)	–	112×112	0.0124	0.8917	0.9125	0.0090	0.8511	0.8952	0.0341	0.8494	0.8671

1) 224×224 PSPNet to 112×112 PSPNet: First of all, we employ PSPNet trained at 224×224 as the teacher and PSPNet trained at 112×112 as the student to validate the knowledge transfer capability of SCRKD. As shown in Table II, our approach achieves the most competitive performance across most metrics on the three datasets compared to a series of state-of-the-art methods. Notably, on ORSSD and EORSSD datasets, the proposed SCRKD improves the F-measure score by over 2% compared to the baseline student, demonstrating impressive performance boosts. Among the competitors, ReviewKD [43] and AT [42] show the strongest advantages, achieving sub-optimal results on several metrics. Additionally, FAMKD [46] and STONet [47] exhibit competitive performance on the ORSI-4199 dataset but fail to comprehensively surpass the proposed SCRKD. As illustrated in Fig. 4, SCRKD overcomes interference from complex backgrounds and cluttered objects in five typical examples, producing predictions that are closest to the ground truths. In summary, SCRKD benefits from comprehensive knowledge transfer at the relation, feature, and saliency prediction levels, leading to significant improvements across all three datasets from 224×224 to 112×112.

2) 224×224 SegFormer to 112×112 SegFormer: To demonstrate the generalizability and model-agnostic nature of SCRKD, we employ SegFormer [11], a Transformer-based architecture, as the baseline and investigate the cross-resolution scenario from 224×224 to 112×112. The experimental results are presented in Table III. Although it does not outperform other methods on the S-measure for the ORSI-4199 dataset, the difference is negligible. It is worth noting that ReviewKD [43] still achieves sub-optimal results in this scenario, reflecting the effectiveness of feature review for cross-resolution distillation. However, ReviewKD solely focuses on feature-level knowledge transfer and fails to explore self-correlation and saliency prediction-level knowledge, which limits its applicability to RSI-SOD tasks. This further underscores the necessity of our proposed SCRKD framework. As can be observed, SCRKD outperforms a series of state-of-the-art methods and achieves performance that is remarkably close to that of the teacher model on the ORSI-4199 dataset. This significantly mitigates the performance degradation caused by insufficient contextual information at low resolutions, providing an effective solution for low-resolution RSI-SOD.

3) 224×224 TransXNet to 112×112 TransXNet: Furthermore, we incorporate a powerful hybrid CNN-Transformer model, TransXNet [12], as the baseline framework, to validate cross-resolution knowledge distillation from 224×224 to 112×112. As evidenced by Table IV, our SCRKD outperforms 10 state-of-the-art distillation methods and 2 lightweight ap-

TABLE III
QUANTITATIVE CROSS-RESOLUTION DISTILLATION PERFORMANCE ON THREE RSI-SOD DATASETS WITH 12 KNOWLEDGE DISTILLATION APPROACHES. THE TOP THREE RESULTS ARE MARKED IN RED, GREEN AND BLUE, RESPECTIVELY.

SegFormer [11]	Publication	Input Size	ORSSD Dataset [14]			EORSSD Dataset [15]			ORSI-4199 Dataset [16]		
			MAE ↓	F_β ↑	S_m ↑	MAE ↓	F_β ↑	S_m ↑	MAE ↓	F_β ↑	S_m ↑
Teacher	NeurIPS 2021	224×224	0.0113	0.9139	0.9273	0.0070	0.8917	0.9235	0.0307	0.8635	0.8765
Student	NeurIPS 2021	112×112	0.0163	0.8681	0.8995	0.0090	0.8459	0.8947	0.0336	0.8499	0.8690
+ FitNet [38]	ICLR 2016	112×112	0.0147	0.8757	0.9019	0.0088	0.8519	0.8973	0.0322	0.8513	0.8714
+ AT [42]	CVPR 2017	112×112	0.0153	0.8726	0.9005	0.0087	0.8509	0.8989	0.0320	0.8534	0.8722
+ CWD [36]	ICCV 2021	112×112	0.0146	0.8743	0.8999	0.0085	0.8539	0.8957	0.0319	0.8537	0.8732
+ ReviewKD [43]	CVPR 2021	112×112	0.0145	0.8781	0.9047	0.0084	0.8585	0.9015	0.0318	0.8540	0.8723
+ SKD [40]	TPAMI 2023	112×112	0.0162	0.8691	0.8996	0.0084	0.8558	0.8994	0.0324	0.8492	0.8699
+ SRD [44]	AAAI 2024	112×112	0.0161	0.8715	0.8967	0.0088	0.8535	0.8974	0.0319	0.8539	0.8718
+ LogitStdKD [45]	CVPR 2024	112×112	0.0148	0.8746	0.9028	0.0088	0.8518	0.8989	0.0317	0.8536	0.8730
+ STONet [47]	TGRS 2024	112×112	0.0164	0.8710	0.9020	0.0090	0.8546	0.9007	0.0332	0.8511	0.8696
+ PD [34]	TPAMI 2024	112×112	0.0166	0.8711	0.8977	0.0091	0.8517	0.8972	0.0320	0.8529	0.8722
+ CMHRD [3]	TGRS 2024	112×112	0.0155	0.8719	0.8987	0.0095	0.8471	0.8966	0.0325	0.8495	0.8686
ISAANet [48]	TGRS 2024	112×112	0.0139	0.8804	0.9108	0.0101	0.8283	0.8861	0.0400	0.8191	0.8495
SOLNet [9]	TGRS 2025	112×112	0.0232	0.8307	0.8718	0.0139	0.8146	0.8710	0.0486	0.8030	0.8260
+ SCRKD (Ours)	–	112×112	0.0141	0.8883	0.9057	0.0079	0.8639	0.9052	0.0310	0.8576	0.8727

TABLE IV
QUANTITATIVE CROSS-RESOLUTION DISTILLATION PERFORMANCE ON THREE RSI-SOD DATASETS WITH 12 KNOWLEDGE DISTILLATION APPROACHES. THE TOP THREE RESULTS ARE MARKED IN RED, GREEN AND BLUE, RESPECTIVELY.

TransXNet [12]	Publication	Input Size	ORSSD Dataset [14]			EORSSD Dataset [15]			ORSI-4199 Dataset [16]		
			MAE ↓	F_β ↑	S_m ↑	MAE ↓	F_β ↑	S_m ↑	MAE ↓	F_β ↑	S_m ↑
Teacher	TNNLS 2025	224×224	0.0102	0.9238	0.9381	0.0064	0.9001	0.9296	0.0293	0.8716	0.8839
Student	TNNLS 2025	112×112	0.0133	0.8799	0.9099	0.0086	0.8471	0.8979	0.0336	0.8492	0.8679
+ FitNet [38]	ICLR 2016	112×112	0.0114	0.8975	0.9232	0.0076	0.8624	0.9107	0.0309	0.8586	0.8771
+ AT [42]	CVPR 2017	112×112	0.0114	0.8993	0.9212	0.0075	0.8620	0.9100	0.0310	0.8581	0.8763
+ CWD [36]	ICCV 2021	112×112	0.0121	0.8984	0.9216	0.0078	0.8645	0.9112	0.0298	0.8600	0.8788
+ ReviewKD [43]	CVPR 2021	112×112	0.0126	0.8918	0.9181	0.0077	0.8652	0.9127	0.0313	0.8593	0.8766
+ SKD [40]	TPAMI 2023	112×112	0.0116	0.8987	0.9216	0.0077	0.8632	0.9086	0.0308	0.8576	0.8763
+ SRD [44]	AAAI 2024	112×112	0.0117	0.8941	0.9194	0.0078	0.8640	0.9089	0.0314	0.8594	0.8759
+ LogitStdKD [45]	CVPR 2024	112×112	0.0121	0.8944	0.9190	0.0081	0.8559	0.9040	0.0305	0.8602	0.8778
+ STONet [47]	TGRS 2024	112×112	0.0109	0.9017	0.9253	0.0076	0.8621	0.9086	0.0305	0.8596	0.8771
+ PD [34]	TPAMI 2024	112×112	0.0125	0.8903	0.9169	0.0076	0.8625	0.9097	0.0305	0.8597	0.8772
+ CMHRD [3]	TGRS 2024	112×112	0.0115	0.9021	0.9248	0.0078	0.8632	0.9087	0.0296	0.8608	0.8793
ISAANet [48]	TGRS 2024	112×112	0.0139	0.8804	0.9108	0.0101	0.8283	0.8861	0.0400	0.8191	0.8495
SOLNet [9]	TGRS 2025	112×112	0.0232	0.8307	0.8718	0.0139	0.8146	0.8710	0.0486	0.8030	0.8260
+ SCRKD (Ours)	–	112×112	0.0100	0.9074	0.9287	0.0069	0.8705	0.9162	0.0295	0.8636	0.8805

proaches across all evaluation metrics on three datasets. Notably, SCRKD enhances the F_β score of 112×112 students by 2.75%, 2.34%, and 1.44% on the three datasets, respectively, validating its consistent generalization capability.

4) 448×448 PSPNet to 224×224 PSPNet: We further investigate the 448×448 to 224×224 distillation scenario using PSPNet as the teacher-student pair. As quantified in Table V, our proposed approach achieves the best performance across all metrics on both datasets. In this cross-resolution setting, CWD [36] and AT [42] produce competitive results, PD [34] and CMHRD [3] also exhibit particular advantages on the EORSSD dataset. By comparison, the proposed SCRKD reaches a 2.04% improvement in F-measure on the ORSSD dataset, which is unmatched by other counterparts. In addition, SCRKD yields 92.95% and 92.70% in terms of S-measure on ORSSD and EORSSD datasets, respectively, showing impressive near-teacher performance. To visualize the predictions of all competitors, we present Fig. 5, which demonstrates that SCRKD can effectively overcome interference from non-salient objects, generating saliency maps with the highest completeness and clearest boundaries in typical scenarios. This further intuitively reveals the effectiveness of our framework.

5) 448×448 PSPNet to 112×112 PSPNet: As reported in Table V, we employ PSPNet trained at 448×448 as the teacher model and PSPNet trained at 112×112 as the student model to evaluate the performance of various methods. Overall, the proposed SCRKD comprehensively outperforms all competitors across all indicators. Particularly, on the EORSSD dataset, SCRKD achieves an impressive F-measure score of nearly 85%. To address the significant challenges posed by the huge resolution gap in one-step distillation, we introduce an assistant model trained at 224×224 and propose a two-stage Cascaded SCRKD framework. Experimental results indicate that the cascaded version further enhances the performance of student, proving its capability to compensate for the contextual information loss at extremely low resolutions and to learn more representative and fine-grained cross-resolution knowledge. Additionally, we visualize the detection results for this cross-

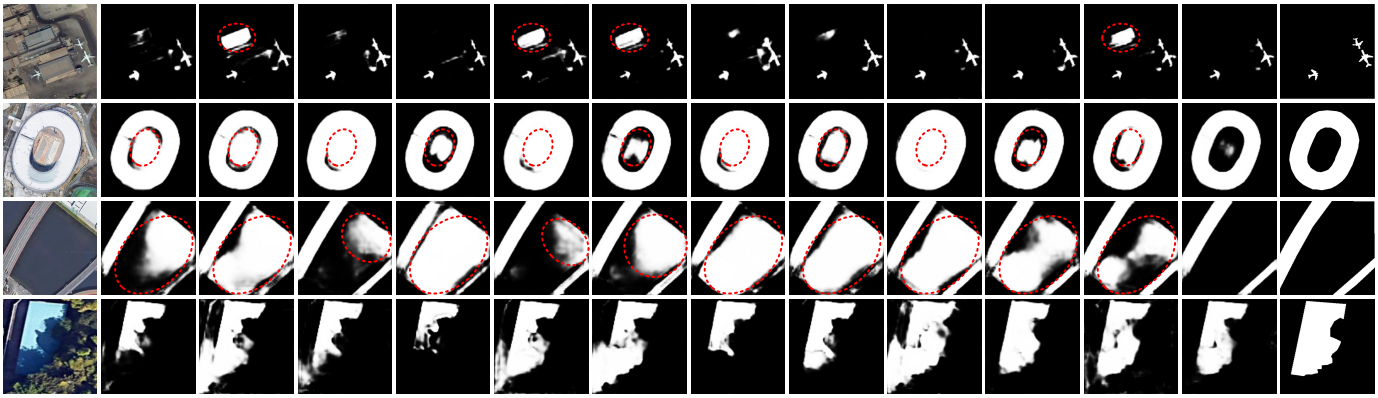


Fig. 4. Saliency map visualization of 224×224 PSPNet to 112×112 PSPNet on the ORSI-4199 dataset. (a) RSIs. (b) Student. (c) FitNet. (d) AT. (e) SKD. (f) CWD. (g) ReviewKD. (h) SRD. (i) LogitStdKD. (j) STONet. (k) PD. (l) CMHRD. (m) SCRKD. (n) GTs.

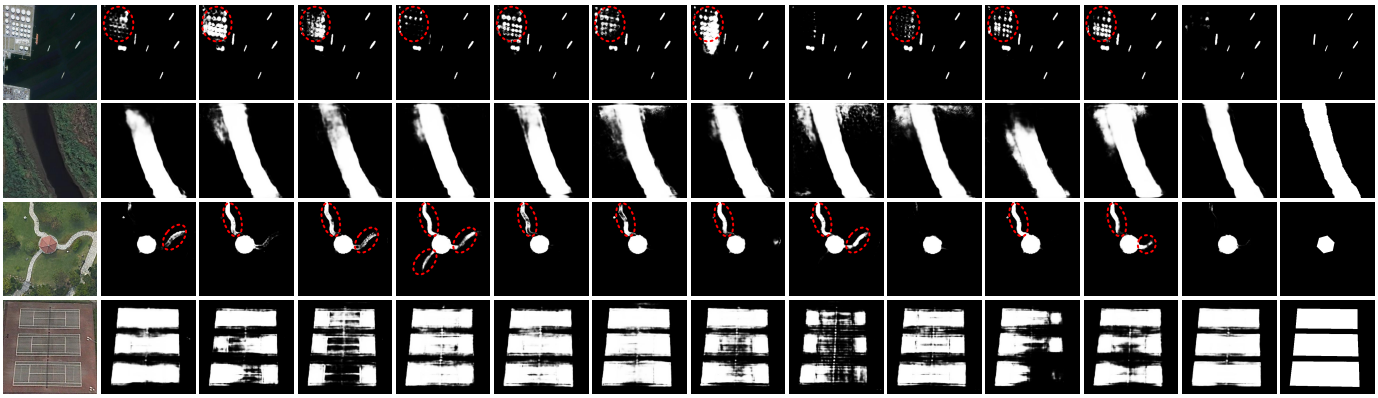


Fig. 5. Saliency map visualization of 448×448 PSPNet to 224×224 PSPNet on the EORSSD dataset. (a) RSIs. (b) Student. (c) FitNet. (d) AT. (e) SKD. (f) CWD. (g) ReviewKD. (h) SRD. (i) LogitStdKD. (j) STONet. (k) PD. (l) CMHRD. (m) SCRKD. (n) GTs.

resolution distillation scenario in Fig. 6. It can be observed that the proposed SCRKD generates saliency maps with clearer boundaries, higher completeness, and fewer false detection, further validating the effectiveness of SCRKD.

6) Comparison of PR and F-measure Curves: Furthermore, we plot the Precision-Recall (PR) and F-measure curves for 12 methods under cross-resolution scenarios on the ORSSD and EORSSD datasets. In the PR curves, the closer the curve is to the top-right corner of the coordinate axis, the better the model's performance. Similarly, in the F-measure curves, the larger the area enclosed by the curve and the coordinate axis, the better the model's performance. As illustrated in Fig. 7, the curves of most comparative methods are intertwined, making it difficult to distinguish their relative performance. In contrast, the red curves representing SCRKD consistently outperforms others in all plots, especially PR curves in Fig. 7(a)-(c) and F-measure curves in Fig. 7(b)-(d), clearly and comprehensively demonstrating the superiority of the proposed framework over various distillation-based competitors.

D. Ablation Study

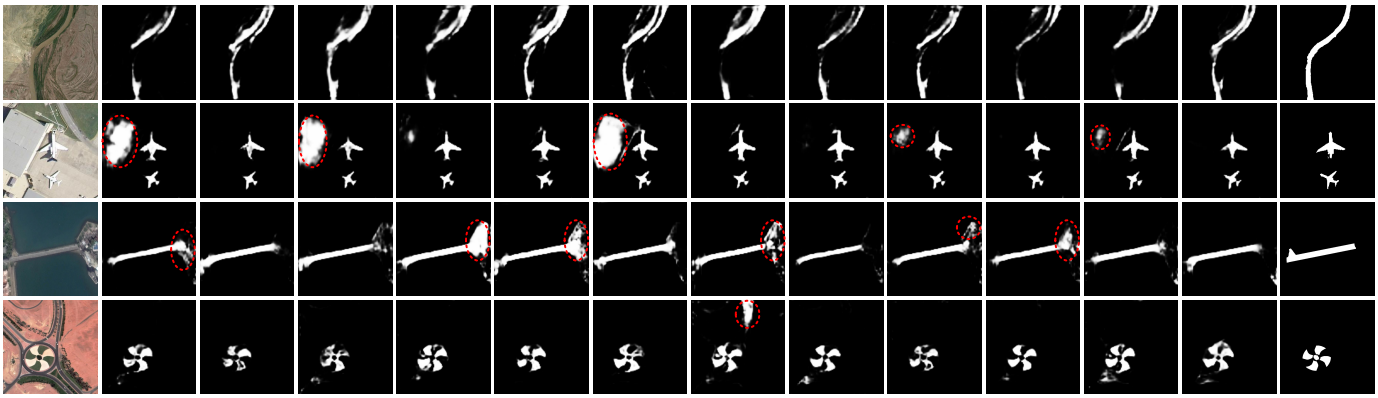
This section presents the ablation study of three distillation modules and impacts of distillation loss coefficients. Specifically, the experimental results under three distillation settings on the EORSSD dataset are presented in Table VI and Fig. 8. Furthermore, Fig. 9 provides 3D visualizations of S-measure under varying coefficient configurations.

1) Quantitative Effects: To address the intrinsic challenges of cross-resolution distillation for RSI-SOD, we present three novel distillation-based modules from multiple perspectives. First, we investigate the individual impacts of three distillation modules on different cross-resolution settings. As shown in Table VI, for various teacher and student models, it can be observed that MSFD contributes the most to the performance boosts of SCRKD across any distillation scenario, followed by MVCD and DSD. Taking the distillation from 224×224 to 112×112 as an example, the introduction of any single distillation module (whether MSFD, MVCD, or DSD) enables the student model to achieve an F-measure score exceeding 84%. This comprehensively validates the effectiveness of our omni-dimensional self-correlation distillation, dynamic similarity-based feature distillation, and decoupled saliency distillation for low-resolution and extremely low-resolution RSI-SOD.

Interestingly, for the same 112×112 student, the introduction of teachers at different resolutions leads to varying performance improvements. When a more powerful 448×448 teacher is employed, compared to a 224×224 teacher, both MVCD and MSFD provide greater performance gains to

TABLE V
QUANTITATIVE CROSS-RESOLUTION DISTILLATION PERFORMANCE ON THREE RSI-SOD DATASETS WITH 12 KNOWLEDGE DISTILLATION APPROACHES. THE TOP THREE RESULTS ARE MARKED IN RED, GREEN AND BLUE, RESPECTIVELY.

PSPNet [10]	Publication	**448×448 → 224× 224**						**448×448 → 112× 112**					
		ORSSD Dataset [14]			EORSSD Dataset [15]			ORSSD Dataset [14]			EORSSD Dataset [15]		
		MAE↓	F_β↑	S_m↑	MAE↓	F_β↑	S_m↑	MAE↓	F_β↑	S_m↑	MAE↓	F_β↑	S_m↑
Teacher	CVPR 2017	0.0104	0.9251	0.9309	0.0071	0.9030	0.9301	0.0104	0.9251	0.9309	0.0071	0.9030	0.9301
Student	CVPR 2017	0.0119	0.8981	0.9146	0.0077	0.8796	0.9135	0.0149	0.8687	0.8944	0.0098	0.8284	0.8796
+ FitNet [38]	ICLR 2016	0.0109	0.9125	0.9231	0.0069	0.8858	0.9212	0.0134	0.8823	0.9021	0.0088	0.8376	0.8887
+ AT [42]	CVPR 2017	0.0100	0.9148	0.9292	0.0072	0.8914	0.9211	0.0134	0.8814	0.9008	0.0096	0.8394	0.8885
+ CWD [36]	ICCV 2021	0.0117	0.9038	0.9229	0.0071	0.8914	0.9216	0.0133	0.8796	0.9047	0.0092	0.8387	0.8891
+ ReviewKD [43]	CVPR 2021	0.0106	0.9117	0.9240	0.0075	0.8888	0.9232	0.0128	0.8836	0.9067	0.0095	0.8401	0.8874
+ SKD [40]	TPAMI 2023	0.0109	0.9049	0.9189	0.0071	0.8840	0.9202	0.0132	0.8781	0.9011	0.0096	0.8325	0.8837
+ SRD [44]	AAAI 2024	0.0113	0.9065	0.9161	0.0076	0.8847	0.9187	0.0149	0.8690	0.8968	0.0096	0.8329	0.8842
+ LogitStdKD [45]	CVPR 2024	0.0130	0.9008	0.9170	0.0077	0.8835	0.9151	0.0138	0.8793	0.8965	0.0089	0.8355	0.8882
+ STONet [47]	TGRS 2024	0.0127	0.9091	0.9231	0.0074	0.8905	0.9199	0.0130	0.8861	0.9038	0.0092	0.8378	0.8872
+ PD [34]	TPAMI 2024	0.0110	0.9069	0.9187	0.0074	0.8873	0.9225	0.0137	0.8829	0.9024	0.0094	0.8360	0.8847
+ CMHRD [3]	TGRS 2024	0.0110	0.9063	0.9230	0.0069	0.8854	0.9200	0.0143	0.8743	0.9007	0.0091	0.8373	0.8853
ISAANet [48]	TGRS 2024	0.0117	0.9054	0.9247	0.0082	0.8770	0.9160	0.0139	0.8804	0.9108	0.0101	0.8283	0.8861
SOLNet [9]	TGRS 2025	0.0243	0.8309	0.8755	0.0125	0.8418	0.8908	0.0232	0.8307	0.8718	0.0139	0.8146	0.8710
+ SCRKD (Ours)	–	0.0096	0.9185	0.9295	0.0064	0.8949	0.9270	0.0119	0.8882	0.9093	0.0088	0.8494	0.8965
+ Cascaded SCRKD	–	–	–	–	–	–	–	0.0117	0.8922	0.9146	0.0083	0.8523	0.8982



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m) (n)

Fig. 6. Saliency map visualization of 448×448 PSPNet to 112×112 PSPNet on the EORSSD dataset. (a) RSIs. (b) Student. (c) FitNet. (d) AT. (e) SKD. (f) CWD. (g) ReviewKD. (h) SRD. (i) LogitStdKD. (j) STONet. (k) PD. (l) CMHRD. (m) SCRKD. (n) GTs.

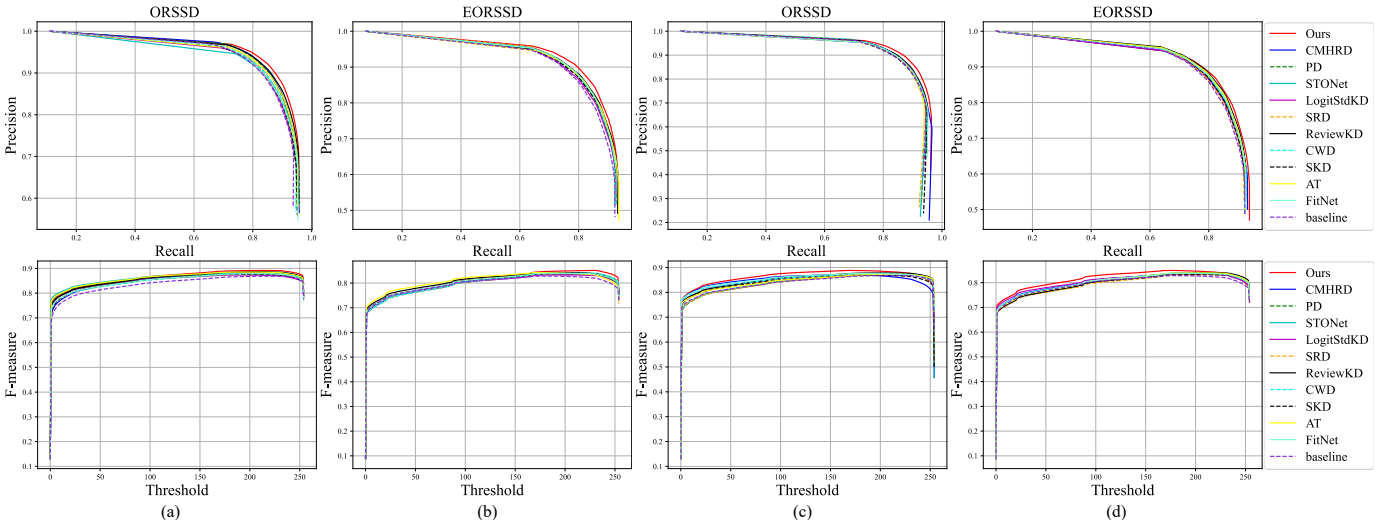
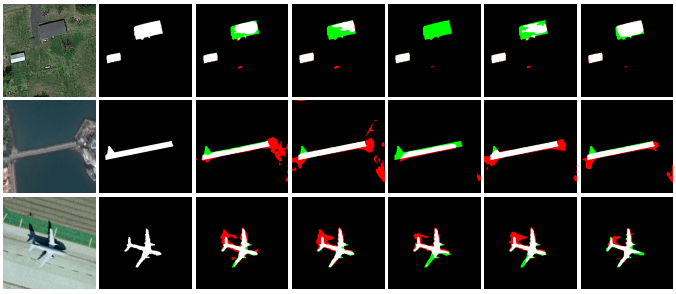


Fig. 7. PR and F-measure curves visualization of 10 state-of-the-art methods. (a)-(b) 224×224 PSPNet to 112×112 PSPNet on the ORSSD and EORSSD datasets. (c) 224×224 SegFormer to 112×112 SegFormer on the ORSSD dataset. (d) 448×448 PSPNet to 112×112 PSPNet on the EORSSD dataset.

TABLE VI
ABLATION EXPERIMENTS OF PSPNET ON THE EORSSD DATASET.

No.	Resolution	MVCD	MSFD	DSD	$F_\beta\uparrow$	$S_m\uparrow$
Distilling from 224×224 teacher to 112×112 student						
1	112×112				0.8284	0.8796
2	112×112	✓			0.8417	0.8932
3	112×112		✓		0.8434	0.8936
4	112×112			✓	0.8401	0.8920
5	112×112	✓	✓		0.8460	0.8941
6	112×112	✓	✓	✓	**0.8511**	**0.8952**
Distilling from 448×448 teacher to 224×224 student						
1	224×224				0.8796	0.9135
2	224×224	✓			0.8866	0.9221
3	224×224		✓		0.8906	0.9232
4	224×224			✓	0.8853	0.9175
5	224×224	✓	✓		0.8920	0.9251
6	224×224	✓	✓	✓	**0.8949**	**0.9270**
Distilling from 448×448 teacher to 112×112 student						
1	112×112				0.8284	0.8796
2	112×112	✓			0.8429	0.8903
3	112×112		✓		0.8460	0.8949
4	112×112			✓	0.8399	0.8900
5	112×112	✓	✓		0.8475	0.8951
6	112×112	✓	✓	✓	**0.8494**	**0.8965**



(a) (b) (c) (d) (e) (f) (g)

Fig. 8. Typical ablation predictions of 448×448 PSPNet to 112×112 PSPNet on the EORSSD dataset, where false positives and missing parts are marked in red and green, respectively. (a) RSIs. (b) GTs. (c) baseline. (d) +MVCD. (e) +MSFD. (f) +MVCD+MSFD. (g) the fully proposed SCRKD.

the student model. However, when three synergistic distillation modules are combined in Table VI, the student achieves an F-measure of 85.11% under the guidance of the 224×224 teacher, but only 84.94% under the supervision of the 448×448 teacher. These results indicate that a higher-resolution teacher can indeed provide more finer-grained structured knowledge, but the huge resolution gap between 448×448 and 112×112 limits the performance of combining these three distillation modules, preventing the student from learning more effectively from the higher-resolution teacher.

When combining MVCD and MSFD, superior performance is achieved across all three cross-resolution scenarios compared to introducing MVCD or MSFD separately. For instance, a 448×448 teacher model enables the 112×112 student model to achieve a 1.91% gain in the F-measure score in Table VI, demonstrating significant performance boosts. Furthermore, with the addition of DSD to MVCD and MSFD, the distillation potential of SCRKD is further unlocked across various cross-resolution settings. Notably, the 224×224 student achieves an impressive F-measure score of 89.49%, substantially mitigating the performance degradation caused by resolution reduction. These results highlight the effectiveness of jointly integrating the three distillation modules, which continuously enhance the distillation capability of SCRKD to address the inherent challenges of low-resolution RSI-SOD.

2) Qualitative Effects: For a more comprehensive visualization of the effects of different distillation modules, Fig. 8 illustrates representative predicted saliency maps from various ablation variants. The results indicate that the student model, when equipped solely with either MVCD or MSFD, fails to adequately address the issues of missed and false detection stemming from insufficient low-resolution contextual information. In contrast, the synergistic integration of MVCD and MSFD, complemented by the full SCRKD model, markedly diminishes the occurrence of missed and false detection pixels, thereby producing saliency maps that exhibit a higher degree of congruence with the ground truth annotations.

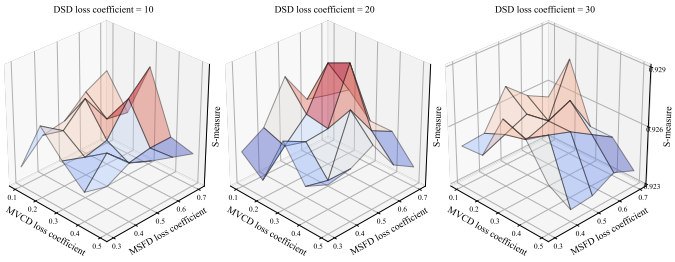


Fig. 9. The variation of S-measure scores with respect to different hyperparameters on the ORSSD dataset for TransXNet-based SCRKD.

3) Impacts of Loss Coefficients: During the training phase of SCRKD, the three distillation losses are jointly optimized with $\mathcal{L}_{SOD}$. Following multi-task learning, we empirically determine the balancing coefficients by considering their magnitudes relative to $\mathcal{L}_{SOD}$, while ensuring $\mathcal{L}_{SOD}$ contributes dominantly to gradient optimization. To investigate the sensitivity of overall performance to different hyperparameter configurations, we conduct a case study on the ORSSD dataset using TransXNet. As clearly shown in Fig. 9, the S-measure exhibits varying degrees of degradation among three surfaces when α or β is set either too high or too low. Notably, the curve for $\gamma = 20$ achieves higher peak performance compared to $\gamma = 10$ and $\gamma = 30$, indicating the existence of optimal value ranges for all three hyperparameters. Importantly, even the lowest S-measure score in Fig. 9 remains above 92.3% and maintains competitiveness with the results in Table IV, demonstrating the robustness of our proposed distillation losses under varying weight configurations.

E. Visualization Analysis

Here, to further elucidate the reasons behind the remarkable performance gains of the proposed SCRKD framework for cross-resolution RSI-SOD, we perform a comprehensive visualization analysis of the presented MVCD, MSFD, and DSD components to provide deeper insights. The typical visualization results are illustrated in Figs. 10, 11, and 12.

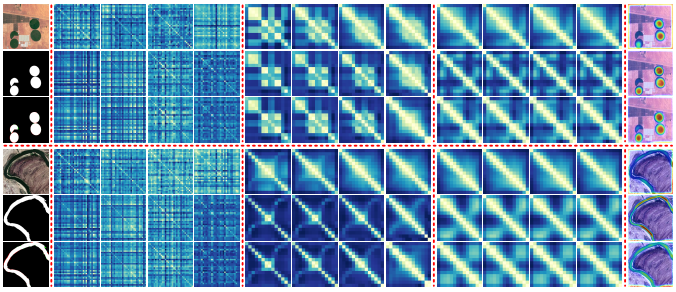


Fig. 10. Multi-view correlation map visualization on the EORSSD dataset. (a) RSIs or GTs or error maps; (b)-(e) channel-wise correlation maps for four stages; (f)-(i) height-wise correlation maps for four stages; (j)-(m) width-wise correlation maps for four stages; (n) spatial feature maps. The first three and last three lines show two individual samples. Each group of three lines represents 224×224 student, 448×448 teacher, and the proposed SCRKD.

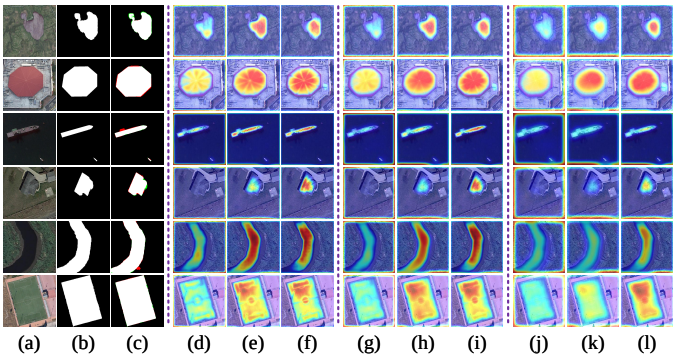


Fig. 11. Typical feature visualization of 448×448 PSPNet to 224×224 PSPNet with our SCRKD on the EORSSD dataset. (a) RSIs. (b) GTs. (c) Error maps. (d)-(f) Stage1 features of student, our SCRKD, and teacher, respectively. (g)-(i) Stage2 features of student, our SCRKD, and teacher, respectively. (j)-(l) Stage3 features of student, our SCRKD, and teacher, respectively.

1) Correlation Map Visualization: As illustrated in Fig. 10, we present the correlation matrices of the student, teacher, and SCRKD across multiple stages in channel-wise, height-wise, and width-wise dimensions for two representative RSIs. Our analysis reveals three key observations: **First**, the self-correlation matrices of both teacher and student models exhibit distinct characteristics across different stages and dimensions for the same sample in Fig. 10(b)-(m), while demonstrating varying patterns across different samples and hierarchical levels. These multi-view self-correlation matrices encapsulate crucial knowledge highly relevant to RSI-SOD, representing one of the significant manifestations of performance variations caused by resolution differences. **Second**, to effectively transfer this valuable knowledge from the high-resolution teacher to the low-resolution student, we develop the MVCD component. The implementation of MVCD induces a notable transformation in the student model's correlation matrices, which progressively align with the characteristic patterns of the teacher. This alignment, which we consider a fundamental factor in performance enhancement, is clearly observable in our ablation analysis. **Third**, after applying MVCD distillation, the student model's feature representations reveal intuitively enhanced activation for salient objects compared to its pre-distillation state in Fig. 10(n). This improved activation provides critical clues for more accurate detection of remote sensing salient objects, as evidenced by our qualitative results.

2) Feature Visualization: To demonstrate the differences in feature representations among the high-resolution teacher model, low-resolution student model, and SCRKD at various stages, we visualize the spatial features of six samples across three stages, as shown in Fig. 11. By comparison, we observe that the teacher, benefiting from higher-resolution contextual inputs, generates the most precise spatial activations for salient objects. In contrast, the student, constrained by resolution limitations, produces incomplete, less accurate, and less prominent activations for salient objects. To address this deficiency, we meticulously design a feature-level distillation technique MSFD. After incorporating MSFD, the spatial features of the student model exhibit closer alignment with those of the teacher model, and the spatial activations for salient objects are significantly enhanced. This improvement definitely creates favorable conditions for boosting detection performance, visually demonstrating the positive contribution of MSFD.

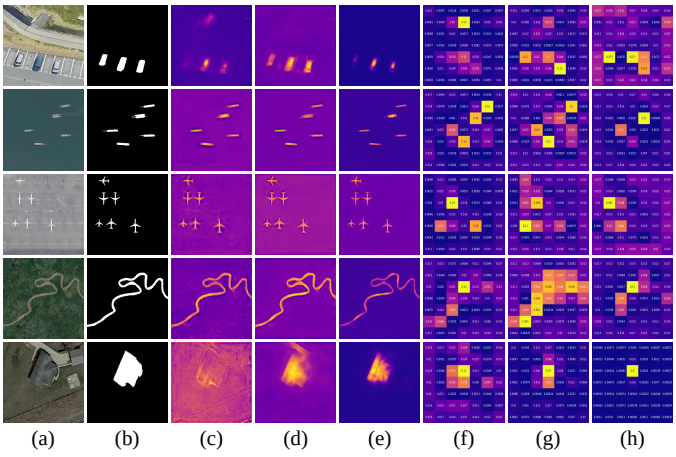


Fig. 12. Typical logits visualization of 448×448 PSPNet to 224×224 PSPNet with SCRKD on the EORSSD dataset. (a) RSIs. (b) GTs. (c)-(e) Stage1 logits heatmaps for student, SCRKD, and teacher, respectively. (f)-(h) Stage5 soft logits heatmaps when $\tau_2 = 4$ for student, SCRKD, and teacher, respectively.

3) Logits Visualization: Finally, we also visualize the logit maps of the student, teacher, and DSD as shown in Fig. 12, which represent the predicted saliency maps before the sigmoid function. For Fig. 12(c)-(e), it can be observed that the student struggles to suppress noise caused by complicated backgrounds and exhibits weaker attention to salient regions. In contrast, the teacher accurately and completely delineates the detailed textures and contours of salient objects. After integrating DSD, the student shows enhanced focus on salient objects, thereby unlocking the distillation potential of SCRKD and improving the detection capability. For the soft logit maps of the final stage, as shown in Fig. 12(f)-(h), we find that while the teacher excels at detecting salient objects, it fails to provide soft labels that are helpful for logit distillation. To address this challenge, we propose DSD, which decouple the single-channel saliency logit map into two-channel category logit map. Our findings indicate that incorporating DSD enables the student to learn more reasonable soft logits that better characterize salient and non-salient regions, thereby

facilitating the student to effectively learn the dark knowledge of saliency prediction level from high-resolution teachers.

V. Conclusion

In this work, we propose a cross-resolution distillation framework SCRKD for RSI-SOD. To further boost the performance for low-resolution models, we extend vanilla SCRKD to Cascaded SCRKD. In our experiments, we find several interesting insights as follows: 1) Under fixed teacher-student resolution settings, a more powerful teacher (e.g., TransXNet vs. PSPNet and SegFormer) can yield a stronger student. 2) For a student with a fixed resolution, its performance degrades as the resolution gap with the teacher increases, due to the larger domain gap induced by higher-resolution teachers. 3) When the resolution gap between student and teacher models exceeds $2\times$, our proposed Cascaded SCRKD can produce significantly stronger student models compared to vanilla SCRKD with arbitrary-resolution teachers. In future work, we aim to extend the cross-resolution distillation framework to other remote sensing dense prediction tasks.

Acknowledgment

The numerical calculations in this work have been supported by the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 270–294, 2022.

[2] Y. Liu, Z. Xiong, Y. Yuan, and Q. Wang, "Distilling knowledge from super-resolution for efficient remote sensing salient object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023, Art no. 5609116.

[3] Y. Zhang, X. Lei, Q. Hu, C. Xu, W. Yang, and G.-S. Xia, "Learning cross-modality high-resolution representation for thermal small-object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024, Art no. 5404815.

[4] Y. Liu, Z. Xiong, Y. Yuan, and Q. Wang, "Transcending pixels: Boosting saliency detection via scene understanding from aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023, Art no. 5616416.

[5] M.-M. Cheng, S.-H. Gao, A. Borji, Y.-Q. Tan, Z. Lin, and M. Wang, "A highly efficient model to study the semantics of salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8006–8021, 2022.

[6] Z. Wang, Y. Zhang, Y. Liu, C. Qin, S. A. Coleman, and D. Kerr, "Larnet: Towards lightweight, accurate and real-time salient object detection," *IEEE Trans. Multimedia*, vol. 26, pp. 5207–5222, 2024.

[7] G. Li, Z. Liu, X. Zhang, and W. Lin, "Lightweight salient object detection in optical remote-sensing images via semantic matching and edge alignment," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, 2023, Art no. 5601111.

[8] G. Li, Z. Liu, Z. Bai, W. Lin, and H. Ling, "Lightweight salient object detection in optical remote sensing images via feature correlation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022, Art no. 5617712.

[9] Z. Li, Y. Miao, X. Li, W. Li, J. Cao, Q. Hao, D. Li, and Y. Sheng, "Speed-oriented lightweight salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–14, 2025, Art no. 5601014.

[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 6230–6239.

[11] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2021, pp. 12 077–12 090.

[12] M. Lou, S. Zhang, H.-Y. Zhou, S. Yang, C. Wu, and Y. Yu, "Transxnet: Learning both global and local dynamics with a dual dynamic token mixer for visual recognition," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2025.

[13] Z. Xiong, Y. Liu, Q. Wang, and X. X. Zhu, "Rssod-bench: A large-scale benchmark dataset for salient object detection in optical remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2023, pp. 6549–6552.

[14] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, "Nested network with two-stream pyramid for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9156–9166, 2019.

[15] Q. Zhang, R. Cong, C. Li, M.-M. Cheng, Y. Fang, X. Cao, Y. Zhao, and S. Kwong, "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Trans. Image Process.*, vol. 30, pp. 1305–1317, 2021.

[16] Z. Tu, C. Wang, C. Li, M. Fan, H. Zhao, and B. Luo, "Orsi salient object detection via multiscale joint region and boundary model," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, Art no. 5607913.

[17] Y. Liu, Y. Yuan, and Q. Wang, "Uncertainty-aware graph reasoning with global collaborative learning for remote sensing salient object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[18] Y. Xie, S. Liu, H. Chen, S. Cao, H. Zhang, D. Feng, Q. Wan, J. Zhu, and Q. Zhu, "Localization, balance, and affinity: A stronger multifaceted collaborative salient object detector in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–17, 2025, Art no. 4700117.

[19] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[20] R. Yan, L. Yan, G. Geng, Y. Cao, P. Zhou, and Y. Meng, "Asnet: Adaptive semantic network based on transformer–cnn for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024, Art no. 5608716.

[21] K. Liu, B. Zhang, J. Lu, and H. Yan, "Toward integrity and detail with ensemble learning for salient object detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024, Art no. 5624615.

[22] Y. Liu and L. Zhang, "Multimodal decomposed distillation with instance alignment and uncertainty compensation for thermal object detection," in *Proc. 33rd ACM Int. Conf. Multimedia (ACM MM)*, Oct. 2025, pp. 1–10.

[23] Z. Dong, G. Gao, T. Liu, Y. Gu, and X. Zhang, "Distilling segmenters from cnns and transformers for remote sensing images' semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023, Art no. 5613814.

[24] C. Pang, X. Weng, J. Wu, Q. Wang, and G.-S. Xia, "Hicd: Change detection in quality-varied images via hierarchical correlation distillation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024, Art no. 5611816.

[25] J. Wu, L. Fang, and J. Yue, "Takd: Target-aware knowledge distillation for remote sensing scene classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 9, pp. 8188–8200, 2024.

[26] G. Xu, X. Jiang, Y. Zhou, S. Li, X. Liu, and P. Lin, "Robust land cover classification with multimodal knowledge distillation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.

[27] C. Li, G. Cheng, G. Wang, P. Zhou, and J. Han, "Instance-aware distillation for efficient object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, 2023.

[28] Z. Feng, J. Lai, and X. Xie, "Resolution-aware knowledge distillation for efficient inference," *IEEE Trans. Image Process.*, vol. 30, pp. 6985–6996, 2021.

[29] S. Ge, S. Zhao, C. Li, Y. Zhang, and J. Li, "Efficient low-resolution face recognition via bridge distillation," *IEEE Trans. Image Process.*, vol. 29, pp. 6898–6908, 2020.

[30] K. Zhang, S. Ge, R. Shi, and D. Zeng, "Low-resolution object recognition with cross-resolution relational contrastive distillation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2374–2384, 2024.

[31] C. Ma, Q. Guo, Y. Jiang, P. Luo, Z. Yuan, and X. Qi, "Rethinking resolution in the context of efficient video recognition," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2022, pp. 37 865–37 877.

[32] S. Shin, J. Lee, J. Lee, Y. Yu, and K. Lee, "Teaching where to look: Attention similarity knowledge distillation for low resolution face recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 631–647.

[33] Y. Zhu, Q. Zhou, N. Liu, Z. Xu, Z. Ou, X. Mou, and J. Tang, "Scalekd: Distilling scale-aware knowledge in small object detector," in *Proc. IEEE*

Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2023, pp. 19 723–19 733.
[34] G. Guo, D. Zhang, L. Han, N. Liu, M.-M. Cheng, and J. Han, "Pixel distillation: Cost-flexible distillation across image sizes and heterogeneous networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 9536–9550, 2024.
[35] K. Wang, Z. Wang, Z. Li, X. Teng, and Y. Li, "Multi-scale cross distillation for object detection in aerial images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Nov. 2024, pp. 452–471.
[36] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5311–5320.
[37] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1365–1374.
[38] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–9.
[39] L. Qi, J. Kuen, J. Gu, Z. Lin, Y. Wang, Y. Chen, Y. Li, and J. Jia, "Multi-scale aligned distillation for low-resolution detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14 443–14 453.
[40] Y. Liu, C. Shu, J. Wang, and C. Shen, "Structured knowledge distillation for dense prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7035–7049, 2023.
[41] Y. Bai, Y. Liu, and Y. Li, "Learning frequency-aware cross-modal interaction for multimodal fake news detection," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 5, pp. 6568–6579, 2024.
[42] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–13.
[43] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5008–5017.
[44] R. Miles and K. Mikolajczyk, "Understanding the role of the projector in knowledge distillation," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Mar. 2024, pp. 4233–4241.
[45] S. Sun, W. Ren, J. Li, R. Wang, and X. Cao, "Logit standardization in knowledge distillation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 15 731–15 740.
[46] C. Pham, V.-A. Nguyen, T. Le, D. Phung, G. Carneiro, and T.-T. Do, "Frequency attention for knowledge distillation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 2277–2286.
[47] W. Zhou, P. Yang, W. Qiu, and F. Qiang, "Stonet-s*: A knowledge-distilled approach for semantic segmentation in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024, Art no. 4414413.
[48] Z. Yao and W. Gao, "Iterative saliency aggregation and assignment network for efficient salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024, Art no. 5633213.

Yanfeng Liu (Student Member, IEEE) received the M.S. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2024. He is currently pursuing the Ph.D. degree with the School of Computer Science, Wuhan University, Wuhan, China. He received the Outstanding Master's Thesis Nomination Award from China Society of Image and Graphics (CSIG) in 2024 and was recognized as the TOP 50 Reviewer by IEEE GRSL in 2025.

His research interests include computer vision, remote sensing, and multimedia signal processing.

Xin Zhang received the B.E. degree from Chongqing University, Chongqing, China, in 2009, and M.S. degree from University of Electronic Science and Technology of China, Chengdu, China, in 2025. He is currently a senior engineer with the Multisensor Intelligent Detection and Recognition Technologies R&D Center, China Aerospace Science and Technology Corporation (CASC), Chengdu, China.

His research interests include radar signal processing and electronic embedded system development.

Wei Guo received the Ph.D. degree from Chongqing University, Chongqing, China, in 2021. She is currently an engineer with the Multisensor Intelligent Detection and Recognition Technologies R&D Center, China Aerospace Science and Technology Corporation (CASC), Chengdu, China.

Her research interests include radar target recognition, multimodal target recognition, and machine learning.

Lefei Zhang (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Wuhan University, Wuhan, China, in 2008 and 2013, respectively. He was a Big Data Institute Visitor with the Department of Statistical Science, University College London, U.K., and a Hong Kong Scholar with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China. He is a professor with the School of Computer Science, Wuhan University, Wuhan, China, and also with the Hubei Luojia Laboratory, Wuhan, China. His research interests include pattern recognition, image processing, and remote sensing.

Dr. Zhang serves as an Associate Editor for IEEE Transactions on Geoscience and Remote Sensing, and IEEE Geoscience and Remote Sensing Letters.