SRTAR: SUPER-RESOLUTION INJECTION WITH TASK-AWARE REFINEMENT FOR CHALLENGING OBJECT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient extraction of high-quality features from input images is critical for accurate challenging object segmentation. Real-world inputs often contain low-resolution (LR) contents, in this circumstance, even employing transformer architectures providing powerful feature modeling, achieving satisfactory performance is limited by the lack of high-frequency components crucial for the task within LR content. Image super-resolution (SR) is one of the promising solutions. However, due to the ill-posed property of SR, typical SR methods struggle to restore task-relevant high-frequency contents, which may dilute the advantage of utilizing the SR method. Therefore, we propose Super-Resolution Injection with Task-Aware Refinement (SRTAR) that effectively guides the generation of SR images beneficial to achieving satisfactory segmentation performance when processing LR images. The critical component is task-aware refinement (TAR) that enables the SR branch to acquire task-relevant discriminative knowledge from segmentation tasks and provides detail-rich features for segmentation refinement. Moreover, we propose a mixed-quality patch stack and a trilateral fusion loss to enhance the efficacy of TAR by addressing potential problems when employing the TAR. Extensive experiments demonstrate that our SRTAR achieves outstanding performance by generating SR images useful for specific task across 8 challenging object segmentation tasks.

1 INTRODUCTION

Binary object segmentation is a fundamental computer vision task encompassing camouflaged object detection (COD) (Pang et al., 2024), salient object detection (SOD) (Luo et al., 2024), medical image segmentation (Fan et al., 2020b; Wang et al., 2024; Yin et al., 2022; Zhou et al., 2023a), among others. Although their semantics differ, these tasks all require complete object regions with accurate boundaries under challenging foreground and background appearances. In real-world applications such as medical image segmentation and surveillance systems, task networks may encounter low-resolution (LR) inputs because of camera or storage limitations and the presence of tiny objects. As a result, weak boundaries, fine structures, and subtle transitions between foreground and background may be severely degraded, limiting task performance. More recently, Vision Transformers (Wang et al., 2022; Liu et al., 2021), which provide strong feature modeling by integrating self-attention and feedforward networks at each encoding layer, have become mainstream architectures for binary object segmentation. However, enhancing the network architecture alone cannot fully compensate for visual cues degraded at the input level.

A natural strategy is to first enhance the LR inputs with image super resolution (SR) which can restore missing high-frequency details, and then perform segmentation on the reconstructed images. Conventional SR restoration primarily through pixel-wise losses (Haris et al., 2018; Kim et al., 2016; Liang et al., 2021; Lim et al., 2017; Zhang et al., 2017; 2018) or perceptual losses (Bruna et al., 2016; Gatys et al., 2015; Johnson et al., 2016). However, due to the ill-posed nature of SR, these approaches may not ensure the reconstruction of features crucial for achieving better performance for downstream segmentation tasks, thereby leading to marginal performance improvement from SR. As illustrated in Figure 1, a generic SR model substantially improves image quality over simple upsampling, while the corresponding segmentation improvement remains limited compared with the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

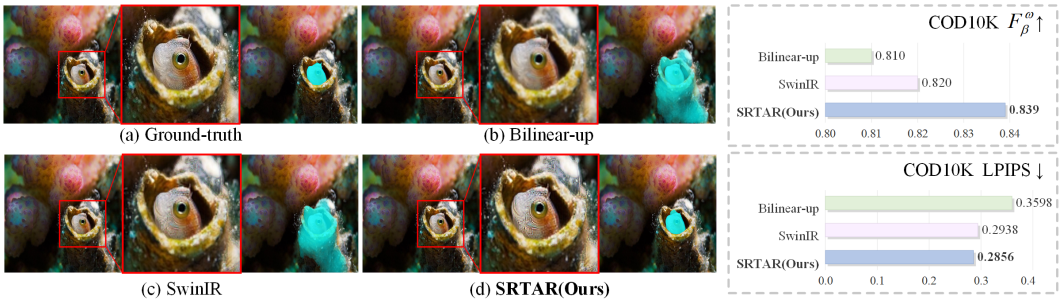


Figure 1: Visualizations and comparisons of the results on COD task with COD10K dataset (Fan et al., 2020a).(a) The ground-truth image and segmentation mask. (b) The bilinear-upsampled image and the predicted result when trained on bilinear-upsampled images. (c) the SR result of the SwinIR (Liang et al., 2021) and the predicted result when trained on the SR images. (d) the SR result and the predicted result when trained by the proposed SRTAR framework. Red boxes indicate enlarged local regions for detailed comparison. The downsampling scale factor is $\times 4$. Additionally, we provide F_β^ω for segmentation performance and LPIPS for SR perceptual quality.

ground truth. This discrepancy indicates that reconstruction quality and downstream task utility are not fully aligned.

To this end, we propose a generic framework for challenging object segmentation, named Super-Resolution Injection with Task-Aware Refinement (SRTAR). Instead of using SR solely as an independent preprocessing stage, SRTAR allows the reconstruction and segmentation branches to interact during decoding so that restored information can be adapted according to downstream tasks. Although reconstructed images provide richer visual cues, repeated downsampling in segmentation encoders may attenuate fine structures and boundary information before decoding. We therefore introduce Task-Aware Refinement (TAR), which injects stage-matched SR content into multiple decoder stages and uses task-relevant gradient guidance to refine segmentation representations. In this way, restored information can be selectively preserved and exploited during hierarchical decoding, improving the utilization of fine-grained cues for accurate foreground segmentation.

To provide TAR with informative local cues at different decoding stages, we further introduce Mixed-Quality Patch Stack (MQPS). We integrate the information from HR and SR images and construct stage-matched patches at multiple spatial scales to provide heterogeneous image-quality cues for hierarchical refinement. These local patches are aligned with different TAR stages, providing scale-appropriate image cues for hierarchical refinement and reducing over-reliance on a single reconstruction quality. In addition, we introduce a Trilateral Fusion Loss (TFL), which constrains the fused SR and segmentation representations from complementary pixel, structural, and gradient perspectives, encouraging more consistent interaction between image reconstruction and segmentation refinement.

We evaluate SRTAR on eight representative binary object segmentation tasks under scale factors of $\times 2$ and $\times 4$. Extensive experiments demonstrate that SRTAR significantly improves task performance while producing perceptually appealing SR results compared with baseline methods, demonstrating its general applicability across binary segmentation tasks. Our contributions are summarized as follows:

- We propose SRTAR, a task-aware SR framework for low-resolution binary object segmentation that jointly considers image reconstruction and segmentation refinement, enabling restored information to be more effectively exploited by downstream segmentation networks.
- We introduce Task-Aware Refinement (TAR), which performs stage-wise SR information injection and task-conditioned gradient refinement. We further develop a stage-matched Mixed-Quality Patch Stack (MQPS) and a Trilateral Fusion Loss (TFL) to provide multi-scale complete local image cues and complementary constraints for SR-segmentation interaction.
- The extensive experiments demonstrate that our SRTAR generally improves performance across eight challenging object segmentation tasks.

2 Related Work

2.1 Binary object segmentation tasks

Binary object segmentation is a fundamental research in computer vision, which aims to achieve precise and complete segmentation of object regions and boundaries from input images by constructing a series of frameworks. It covers diverse tasks, including camouflaged object detection (COD) (Pang et al., 2024; He et al., 2025), salient object detection (SOD) (Duan et al., 2025), polyp segmentation (PS) (Yue et al., 2025), COVID-19 lung infection segmentation (CLI) (Chang et al., 2025), breast lesion segmentation (BLS) (Tang et al., 2023), skin lesion segmentation (SLS) (Wang et al., 2024), transparent object segmentation (TOS) (Fan et al., 2023), and defocus blur detection (DBD) (Zhao et al., 2023), among others. According to their different characteristics, existing methods mainly focus on four directions: visual attention (Kim et al., 2021; Ji et al., 2023), multi-scale feature extraction (Pang et al., 2022; Li et al., 2023), edge refinement (Sun et al., 2022; Zhou et al., 2023b), and integration of different architectures (Zhang et al., 2021; Ji et al., 2021). These approaches have achieved strong performance across various segmentation tasks. However, in real-world scenarios, task-specific architectures may encounter low-resolution (LR) inputs with insufficient visual information. Once discriminative details are lost at the input side, enhancing or reorganizing encoded features alone cannot recover them, which still limits satisfactory segmentation performance. Therefore, restoring task-relevant high-frequency content directly from input images is important for binary object segmentation.

2.2 Image restoration for vision tasks

Recent studies have explored image restoration methods that aim to improve the performance of downstream vision recognition tasks (Cui et al., 2022; Dai et al., 2016; Liu et al., 2018; 2023; Shermeyer & Van Etten, 2019; Zhao et al., 2018b). Wang et al. (Wang et al., 2020) introduce a dual super-resolution learning framework for semantic segmentation to exploit high-resolution representations when processing low-resolution inputs. Haris et al. (Haris et al., 2021) formulate task-driven super-resolution for object detection by jointly optimizing reconstruction and detection objectives. Kim et al. (Kim et al., 2024) propose SR4IR, which introduces a task-driven perceptual loss to transfer task-specific knowledge from a recognition network to the SR network, thereby improving the performance across various high-level vision tasks. Collectively, these studies demonstrate the potential of optimizing image restoration according to the requirements of downstream high-level recognition tasks, beyond the objective of improving image fidelity alone. In contrast to previous approaches, our method focuses on challenging binary segmentation, where weak boundaries, small structures, and subtle foreground background transitions require fine grained spatial refinement. SRTAR injects restored image information into multiple decoding stages and introduces task-aware region gradient refinement to preserve and exploit task relevant details.

3 Proposed method

Overview. Figure 2 shows the proposed SRTAR framework. Given a low-resolution image I_{LR}, SRTAR first reconstructs a super-resolved image I_{SR} using an SR network Liang et al. (2021). The segmentation encoder Liu et al. (2021) subsequently extracts hierarchical features $\{F_i^e\}_{i=1}^4$, and an ASPP module Chen et al. (2018) is employed to the feature F_4^e to form the initial decoder representation. We employ Task Aware Refinement (TAR) to inject stage-matched image cues into multiple decoder stages and employ task-relevant gradient guidance to refine local structure representations. (Section 3.1). Mixed-Quality Patch Stack (MQPS) constructs multi-scale mixed-quality image patches matched to individual TAR stages (Section 3.2), while Trilateral Fusion Loss (TFL) regularizes the interaction between SR and segmentation representations from multi-level perspectives (Section 3.3).

3.1 Task-Aware Refinement

Due to the ill-posed nature of SR, traditional SR reconstruction at the image level does not explicitly guarantee that restored local cues are preserved and effectively utilized throughout hierarchical segmentation network. We therefore introduce Task-Aware Refinement (TAR), which performs two

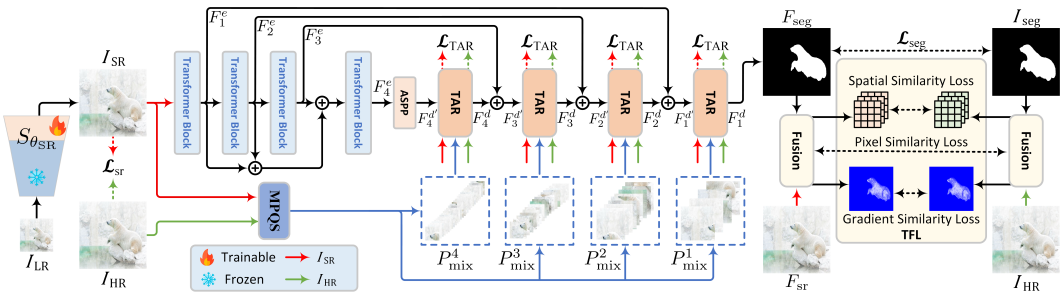


Figure 2: Architecture of the SRTAR. Our SRTAR consists of an SR network, a segmentation network, a series of Task-Aware Refinement (TAR) modules, a Mixed-Quality Patch Stack (MQPS) module, and the Trilateral Fusion Loss (TFL). During inference, the SRTAR relies solely on the reconstructed SR images without HR images.

complementary operations: stage-wise injection of reconstructed information and local structure refinement of task-relevant gradient guidance, as illustrated in Figure 3.

Specifically, the stage-matched image representation P^i_{mix} generated by MQPS is combined with the decoder feature $F^{d'}_i$. The fused feature is processed by parallel deformable convolution branches (Dai et al., 2017) with receptive fields of 1×1, 3×3, and 5×5, together with an average-pooling. The resulting multi-scale features are concatenated and processed by a 1×1 convolution followed by batch normalization and ReLU (BR), yielding the refined representation F^m_i and intermediate predictions $\{M_i\}^3_{i=1}$ under multi-stage supervision. Subsequently, we enable the SR image to mimic the gradient response of the HR image within the task-relevant regions. G^{gt}_h and G^{gt}_s denote the gradient maps extracted from the original HR image I_{HR} and the reconstructed SR image I_{SR}, respectively. The intermediate representation $F^{m'}_i$ is used as a spatial weighting map so that gradient alignment focuses on regions currently considered relevant to the segmentation task. Specifically, $F^{m'}_i$ is multiplied element-wise with gradient maps G^{gt}_h and G^{gt}_s to obtain G^h_i and G^s_i, respectively. By maximizing the similarity between SR and HR gradient responses within task-relevant regions, the TAR supervision is defined as

$$\mathcal{L}_{\text{TAR}} = \sum_i \mathcal{L}_{\text{BCE}}\left(G^{gt}_s \odot F^{m'}_i, G^{gt}_h \odot F^{m'}_i\right), \tag{1}$$

The task-relevant SR gradient response G^{S}_i is further transformed into an attention map by a convolution followed by a sigmoid function, which is used to modulate the refined decoder representation:

$$A^G_i = \sigma(\text{Conv}(G^{\text{S}}_i)), \tag{2}$$

$$F^d_i = F^m_i \odot A^G_i, \tag{3}$$

The output of each TAR stage F^d_i is combined with the corresponding encoder feature to form the input of the next decoder stage:

$$F^{d'}_i = \text{Upsample}\left(F^d_{i+1} + F^e_i\right), \qquad i = 3, 2, 1. \tag{4}$$

Finally, F^d_1 is projected by a 1×1 convolution to obtain the final segmentation map. In this way, TAR therefore uses reconstructed image cues throughout hierarchical decoding, while gradient-based supervision encourages local boundary and structural responses within task-relevant regions during training.

3.2 Mixed-Quality Patch Stack

The effectiveness of TAR is inherently related to the knowledge provided by the task network. For low-level vision tasks, networks are susceptible to shortcut learning (Geirhos et al., 2020), meaning that they may rely on biased or non-intrinsic features in the training data, leading to poor generalization ability.

In particular, learning biased features can pose challenges in employing the TAR, as biased features can provide inaccurate task guidance to hinder segmentation performance. We therefore introduce

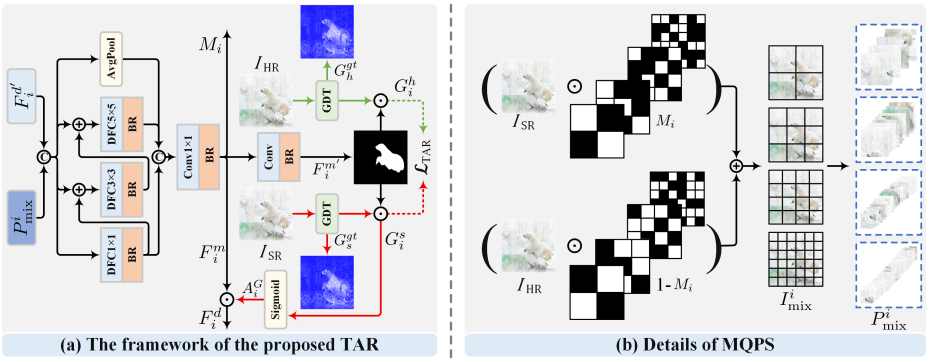


Figure 3: Architecture of (a) TAR and (b) MQPS.

a Stage-Matched Mixed-Quality Patch Stack (MQPS), as illustrated in Figure 3, to provide TAR with heterogeneous image-quality patches and reduce the reliance on biased cues. Inspired by the cross-quality mixing strategy (Kim et al., 2024), MQPS randomly selects either HR or SR content for each grid region to construct a mixed-quality image at each refinement stage:

$$I^i_{\text{mix}} = \mathbf{M}_i \odot I_{\text{SR}} + (1 - \mathbf{M}_i) \odot I_{\text{HR}}. \tag{5}$$

where $\mathbf{M}_i$ is a random binary grid mask with stage-specific grid size that determines whether each spatial region is sampled from the SR or HR image. Different from applying cross-quality mixing strategy solely at the image level, MQPS further constructs the mixed-quality image I^i_{mix} to the patches $P^i_{\text{mix}} = \left\{Patch^N_{k=1}\right\}$ with scale of each decoder stage. By exposing TAR to heterogeneous local reconstruction qualities at multiple decoder scales, MQPS reduces reliance on a single reconstruction-quality pattern and provides stage-appropriate local image cues for refinement. During inference, P^i_{mix} is constructed exclusively from I_{SR} using the same stage-matched cropping procedure.

3.3 Trilateral Fusion Loss

TAR introduces reconstructed image information into hierarchical segmentation decoding, allowing local image cues to participate in task-aware refinement. However, reconstruction and segmentation are optimized toward different objectives, and directly combining their representations may introduce local responses that are poorly aligned with the target regions, especially under complex backgrounds and ambiguous boundaries. We therefore introduce Trilateral Fusion Loss (TFL) to explicitly regularize the interaction between the SR and segmentation branches at the prediction stage, as shown in Figure 2.

Specifically, two 3×3 projection layers are applied to the SR and segmentation branches to extract F_{seg}, $F_{\text{sr}} \in \mathbb{R}^{3\times h\times w}$. We fuse the SR prediction feature F_{sr} with the segmentation prediction feature F_{seg}, and the HR image I_{HR} with the segmentation ground truth label I_{seg}, through element-wise multiplication, forming the fused target-relevant representation F_{fuse} and I_{fuse}, respectively. Then, we constrain the fused representation from three complementary perspectives: pixel consistency, structural consistency, and gradient consistency.

The first constraint focuses on spatial structural consistency between the two fused representations. To reduce the memory cost of direct spatial affinity computation, we apply $16\times$ average pooling to both fused representations to obtain $3 \times 64 \times 64$ matrix and then reshape it to 3×4096. The affinity matrix can be represented as:

$$S^f_{i,j} = \left(\frac{f^{\text{fuse}}_i}{\|f^{\text{fuse}}_i\|_1}\right)^{\text{T}} \left(\frac{f^{\text{fuse}}_j}{\|f^{\text{fuse}}_j\|_1}\right) \tag{6}$$

where $S^f_{i,j}$ denotes the self-correlation between the ith and jth elements of the fused representation F_{fuse}. The corresponding similarity matrix of the GT representation is calculated in the same way, resulting in S^f and S^I as the affinity matrices of f_{fuse} and I_{fuse}, respectively. The spatial similarity loss is defined as

$$\mathcal{L}_{\text{similar}} = \frac{1}{W^2H^2} \sum_{i=1}^{WH} \sum_{j=1}^{WH} \left\| S^f_{i,j} - S^I_{i,j} \right\|_2. \tag{7}$$

where $W = H = 64$. This constraint preserves spatial relationships in task-relevant representations and encourages the segmentation branch to absorb structural information from the SR branch.

We additionally constrain the two fused representations at the pixel level:

$$\mathcal{L}_{\text{consist}} = \frac{1}{C'W'H'} \left\| f^{\text{fuse}} - I^{\text{fuse}} \right\|_1. \tag{8}$$

This term encourages local fused responses to remain consistent with the target-region reference representation.

Finally, we extract gradient responses from the predicted and reference fused representations denoted as f_{gt}^{fuse} and I_{gt}^{fuse} respectively, and introduce the gradient similarity constraint between them to focus the alignment of high-frequency details:

$$\mathcal{L}_{\text{gdt}} = \mathcal{L}_{\text{BCE}}\left(f_{gt}^{fuse}, I_{gt}^{fuse}\right). \tag{9}$$

The complete TFL is formulated as

$$\mathcal{L}_{\text{TFL}} = \mathcal{L}_{\text{similar}} + \mathcal{L}_{\text{consist}} + \mathcal{L}_{\text{gdt}}. \tag{10}$$

In this way, the TFL encourages agreement in target-related appearance, spatial structure, and boundary detail, thereby reducing interference from irrelevant reconstructed content and improving task-aware representation learning.

3.4 Loss Functions

To achieve the joint learning of the whole model during training, we integrate the SR loss, segmentation loss, TAR loss, and TFL loss into a total loss function, so that the entire model can be optimized end-to-end. The final loss function is a weighted combination of these losses:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{seg}} + \lambda_2 \mathcal{L}_{\text{sr}} + \lambda_3 \mathcal{L}_{\text{TAR}} + \lambda_4 \mathcal{L}_{\text{TFL}} \tag{11}$$

where $\mathcal{L}_{\text{seg}}$ combines binary cross-entropy (BCE) loss, intersection over union (IoU) loss , and structural similarity index measure (SSIM) loss, while $\mathcal{L}_{\text{sr}}$ is the pixel-wise reconstruction loss. The weights λ_1, λ_2, λ_3, and λ_4 are set to 1, 0.5, 1, and 1 respectively to balance the task network. The segmentation losse is defined as

$$\mathcal{L}_{\text{seg}} = 30\mathcal{L}_{\text{BCE}} + 0.5\mathcal{L}_{\text{IoU}} + 10\mathcal{L}_{\text{SSIM}}. \tag{12}$$

4 Experiments

4.1 Evaluation

We denote the FPN-style task network and the SR network $S_{\boldsymbol{\theta}_{\text{SR}}}$ as T and S, respectively. Then, we introduce the following training scenarios. $\boldsymbol{I_{\text{LR}}} \rightarrow \boldsymbol{T}$: Train the T with $\mathcal{L}_{\text{seg}}$ using (I_{LR}, y); $\boldsymbol{I_{\text{B}}} \rightarrow \boldsymbol{T}$: Train the T with $\mathcal{L}_{\text{seg}}$ using $\big(B(I_{LR}), y\big)$, where B is bilinear upsampling operator; and $\boldsymbol{I_{\text{S}}} \rightarrow \boldsymbol{T}$: Train the S with $\mathcal{L}_{\text{pixel}}$ using (I_{LR}, I_{HR}) first. Then train T with $\mathcal{L}_{\text{seg}}$ using $\big(S(I_{LR}), y\big)$, while freezing S, as in previous work Zhou et al. (2020). We additionally evaluate SRTAR_{B}, which follows the same task aware training procedure as SRTAR while replacing SwinIR with bilinear upsampling only during inference. This variant evaluates whether the task aware segmentation pathway can retain the benefits learned during training without executing the computationally expensive SR network at deployment time.

4.2 Implementation Details

Following the conventional SR datasets (Agustsson & Timofte, 2017; Bevilacqua et al., 2012), we construct the corresponding LR datasets applying bicubic downsampling with scale factors $\times 4$ and $\times 2$. We adopt two resolution settings: input images are resized to 256×256 or 512×512 and reconstructed to 1024×1024, corresponding to $\times 4$ and $\times 2$ SR, respectively. SwinIR is initialized with DIV2K-pretrained weights (Agustsson & Timofte, 2017), and Swin-L serves as the task encoder.

Table 1: Quantitative comparisons of SRTAR and SRTAR$_{\mathrm{B}}$ with different training scenarios and other SOTAs across eight binary object segmentation tasks, including COD, SOD, PS, CLI, BLS, SLS, TOS, and DBD. Orange-red and teal values denote the best and second-best results.

Method (COD)	Size	CAMO			COD10K			Method (SOD)	Size	DUTS-TE		PASCAL-S	
		$S_m\uparrow$	$F_\beta^\omega\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$F_\beta^\omega\uparrow$	$\mathcal{M}\downarrow$			$S_m\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$\mathcal{M}\downarrow$
ZoomNeXt (Pang et al., 2024)	384	0.888	0.859	0.040	0.898	0.838	**0.017**	VSCode (Luo et al., 2024)	384	0.926	0.024	0.886	0.050
CGCOD (Zhang et al., 2025)	448	0.896	0.864	0.036	0.890	0.824	0.018	VST++ (Liu et al., 2024a)	288	0.923	0.024	0.889	0.046
GBNet (Wang et al., 2026)	704	0.880	0.847	0.044	0.897	0.837	0.019	CamoDiffusion (Sun et al., 2025)	384	0.921	0.023	–	–
F^2HI (Ren et al., 2026)	384	0.897	0.874	0.035	0.892	0.831	0.018	SENet (Hao et al., 2025)	384	0.926	0.022	0.890	0.046
EVPv2 (Liu et al., 2026)	352	0.848	0.786	0.058	0.843	0.746	0.029	EVPv2 (Liu et al., 2026)	352	0.915	0.027	0.879	0.053
$I_{\mathrm{LR}} \to T$	256	0.866	0.802	0.048	0.844	0.728	0.028	$I_{\mathrm{LR}} \to T$	256	0.919	0.026	0.890	0.047
$I_{\mathrm{B}} \to T$	256	0.900	0.854	0.038	0.900	0.810	0.020	$I_{\mathrm{B}} \to T$	256	0.926	0.022	0.888	0.045
$I_{\mathrm{S}} \to T$	256	0.902	0.860	0.037	0.900	0.820	0.019	$I_{\mathrm{S}} \to T$	256	0.925	0.021	0.890	0.043
SRTAR$_{\mathrm{B}}$ (Ours)	256	0.892	0.871	0.033	0.886	0.826	0.018	SRTAR$_{\mathrm{B}}$ (Ours)	256	0.927	0.021	0.891	0.043
SRTAR (Ours)	256	0.896	0.877	**0.032**	0.894	0.839	**0.017**	SRTAR (Ours)	256	0.932	**0.020**	0.893	0.042
$I_{\mathrm{LR}} \to T$	512	0.892	0.841	0.039	0.886	0.800	0.021	$I_{\mathrm{LR}} \to T$	512	0.932	0.023	0.892	0.042
$I_{\mathrm{B}} \to T$	512	0.901	0.864	0.037	**0.910**	0.835	0.018	$I_{\mathrm{B}} \to T$	512	0.933	0.022	0.894	**0.041**
$I_{\mathrm{S}} \to T$	512	0.902	0.861	0.035	0.906	0.835	0.018	$I_{\mathrm{S}} \to T$	512	**0.935**	0.021	**0.895**	0.044
SRTAR$_{\mathrm{B}}$ (Ours)	512	**0.903**	**0.888**	**0.030**	0.908	**0.863**	**0.015**	SRTAR$_{\mathrm{B}}$ (Ours)	512	**0.936**	**0.020**	**0.895**	**0.041**
SRTAR (Ours)	512	**0.904**	**0.889**	**0.030**	**0.912**	**0.871**	**0.015**	SRTAR (Ours)	512	**0.936**	**0.019**	**0.898**	**0.040**

Method (PS)	Size	Kvasir		Method (CLI)	Size	COVID-19		Method (BLS)	Size	BUSI	
		Dice↑	IoU↑			Dice↑	mIoU↑			Dice↑	IoU↑
Rolling-UNet (Liu et al., 2024b)	256	0.839	0.761	SegGPT (Wang et al., 2023)	448	0.131	0.553	Rolling-UNet (Liu et al., 2024b)	256	0.808	0.716
EMCAD-B2 (Rahman et al., 2024)	352	0.859	0.786	Rolling-UNet (Liu et al., 2024b)	384	0.628	0.764	EMCAD-B2 (Rahman et al., 2024)	256	0.806	0.720
PRBENet (Yue et al., 2025)	384	0.913	0.866	Spider (Zhao et al., 2024)	384	0.696	0.813	DSC-ICL (Gao et al., 2025)	128	0.719	–
PMRNet (Kang, 2026)	–	0.877	0.812	SR-ICL (Chang et al., 2025)	384	0.824	0.872	CMU Net (Tang et al., 2023)	256	0.8416	0.7327
GBNet (Wang et al., 2026)	704	0.905	0.856	M^2SNet (Zhao et al., 2026)	352	0.702	0.818	PMRNet (Kang, 2026)	–	0.819	0.736
$I_{\mathrm{LR}} \to T$	256	0.872	0.801	$I_{\mathrm{LR}} \to T$	256	0.218	0.718	$I_{\mathrm{LR}} \to T$	256	0.797	0.730
$I_{\mathrm{B}} \to T$	256	0.886	0.836	$I_{\mathrm{B}} \to T$	256	0.414	0.720	$I_{\mathrm{B}} \to T$	256	0.800	0.764
$I_{\mathrm{S}} \to T$	256	0.897	0.852	$I_{\mathrm{S}} \to T$	256	0.503	0.725	$I_{\mathrm{S}} \to T$	256	0.813	0.780
SRTAR$_{\mathrm{B}}$ (Ours)	256	0.911	0.861	SRTAR$_{\mathrm{B}}$ (Ours)	256	0.820	0.876	SRTAR$_{\mathrm{B}}$ (Ours)	256	0.938	0.894
SRTAR (Ours)	256	0.911	0.860	SRTAR (Ours)	256	0.823	**0.878**	SRTAR (Ours)	256	0.945	0.904
$I_{\mathrm{LR}} \to T$	512	0.889	0.840	$I_{\mathrm{LR}} \to T$	512	0.358	0.767	$I_{\mathrm{LR}} \to T$	512	0.853	0.833
$I_{\mathrm{B}} \to T$	512	0.903	0.858	$I_{\mathrm{B}} \to T$	512	0.477	0.775	$I_{\mathrm{B}} \to T$	512	0.886	0.820
$I_{\mathrm{S}} \to T$	512	0.907	0.862	$I_{\mathrm{S}} \to T$	512	0.512	0.783	$I_{\mathrm{S}} \to T$	512	0.920	0.877
SRTAR$_{\mathrm{B}}$ (Ours)	512	**0.917**	**0.871**	SRTAR$_{\mathrm{B}}$ (Ours)	512	**0.827**	**0.878**	SRTAR$_{\mathrm{B}}$ (Ours)	512	**0.959**	**0.928**
SRTAR (Ours)	512	**0.920**	**0.872**	SRTAR (Ours)	512	**0.829**	**0.880**	SRTAR (Ours)	512	**0.961**	**0.933**

Method (SLS)	Size	ISIC2018		Method (TOS)	Size	Trans10K		Method (DBD)	Size	DUT	
		Dice↑	IoU↑			IoU↑	$\mathcal{M}\downarrow$			$F_\beta\uparrow$	$\mathcal{M}\downarrow$
EMCAD-B2 (Rahman et al., 2024)	352	0.887	0.818	FSPNet (Huang et al., 2023)	384	0.896	0.043	DAD (Zhao et al., 2021b)	320	0.794	0.153
LSSNet (Wang et al., 2024)	352	0.886	0.824	RFENet (Fan et al., 2023)	512	0.892	0.043	EFENet (Zhao et al., 2021a)	320	0.854	0.094
DSC-ICL (Gao et al., 2025)	128	0.815	–	FSEL (Sun et al., 2024)	416	0.892	0.043	MLDBD (Zhao et al., 2023)	320	0.875	0.111
FSEL (Sun et al., 2024)	416	0.885	0.821	CMCM (Lin et al., 2025)	384	0.899	0.046	TCRL (Chai et al., 2024)	320	0.885	0.082
PMRNet (Kang, 2026)	–	0.893	0.825	MirrorSAM (Zha et al., 2026)	512	0.903	0.042	EVP$_{v2}$ (Liu et al., 2026)	320	0.887	0.070
$I_{\mathrm{LR}} \to T$	256	0.889	0.821	$I_{\mathrm{LR}} \to T$	256	0.910	0.044	$I_{\mathrm{LR}} \to T$	256	0.892	0.060
$I_{\mathrm{B}} \to T$	256	0.893	0.823	$I_{\mathrm{B}} \to T$	256	0.942	0.041	$I_{\mathrm{B}} \to T$	256	0.902	0.058
$I_{\mathrm{S}} \to T$	256	0.885	0.827	$I_{\mathrm{S}} \to T$	256	0.942	0.038	$I_{\mathrm{S}} \to T$	256	0.905	0.058
SRTAR$_{\mathrm{B}}$ (Ours)	256	0.894	0.827	SRTAR$_{\mathrm{B}}$ (Ours)	256	0.957	0.021	SRTAR$_{\mathrm{B}}$ (Ours)	256	0.907	0.054
SRTAR (Ours)	256	0.895	0.828	SRTAR (Ours)	256	0.958	0.022	SRTAR (Ours)	256	0.909	0.054
$I_{\mathrm{LR}} \to T$	512	0.890	0.824	$I_{\mathrm{LR}} \to T$	512	0.934	0.038	$I_{\mathrm{LR}} \to T$	512	0.902	0.055
$I_{\mathrm{B}} \to T$	512	0.893	0.822	$I_{\mathrm{B}} \to T$	512	0.951	0.032	$I_{\mathrm{B}} \to T$	512	0.906	0.055
$I_{\mathrm{S}} \to T$	512	**0.896**	0.827	$I_{\mathrm{S}} \to T$	512	0.948	0.031	$I_{\mathrm{S}} \to T$	512	0.908	**0.053**
SRTAR$_{\mathrm{B}}$ (Ours)	512	0.895	**0.830**	SRTAR$_{\mathrm{B}}$ (Ours)	512	**0.962**	**0.019**	SRTAR$_{\mathrm{B}}$ (Ours)	512	**0.914**	**0.052**
SRTAR (Ours)	512	**0.898**	**0.831**	SRTAR (Ours)	512	**0.961**	**0.020**	SRTAR (Ours)	512	**0.915**	**0.052**

Only the last residual Swin Transformer group of SwinIR is updated during training, while the remaining SR blocks are frozen. Each model is trained for 150 epochs using AdamW (Loshchilov & Hutter, 2019), with initial learning rates of 10^{-6} for trainable SR parameters and 10^{-5} for all other trainable parameters. All experiments are implemented in PyTorch (Paszke et al., 2019) on two NVIDIA RTX 5000 GPUs with batch size of $N = 2$ per GPU.

4.3 Quantitative Results

We evaluate eight binary segmentation tasks against SOTA methods using standard metrics, with dataset and metric details provided in Appendix B. Table 1 compare native LR training ($I_{\mathrm{LR}} \to T$), bilinear enlargement ($I_{\mathrm{B}} \to T$), sequential SR preprocessing ($I_{\mathrm{S}} \to T$), SRTAR$_{\mathrm{B}}$, and SRTAR. Bilinear enlargement generally improves native LR training, confirming that insufficient spatial resolution limits segmentation performance. Sequential SR further improves most metrics, but its gains remain limited on some benchmarks because pixel-oriented reconstruction does not explicitly pre-

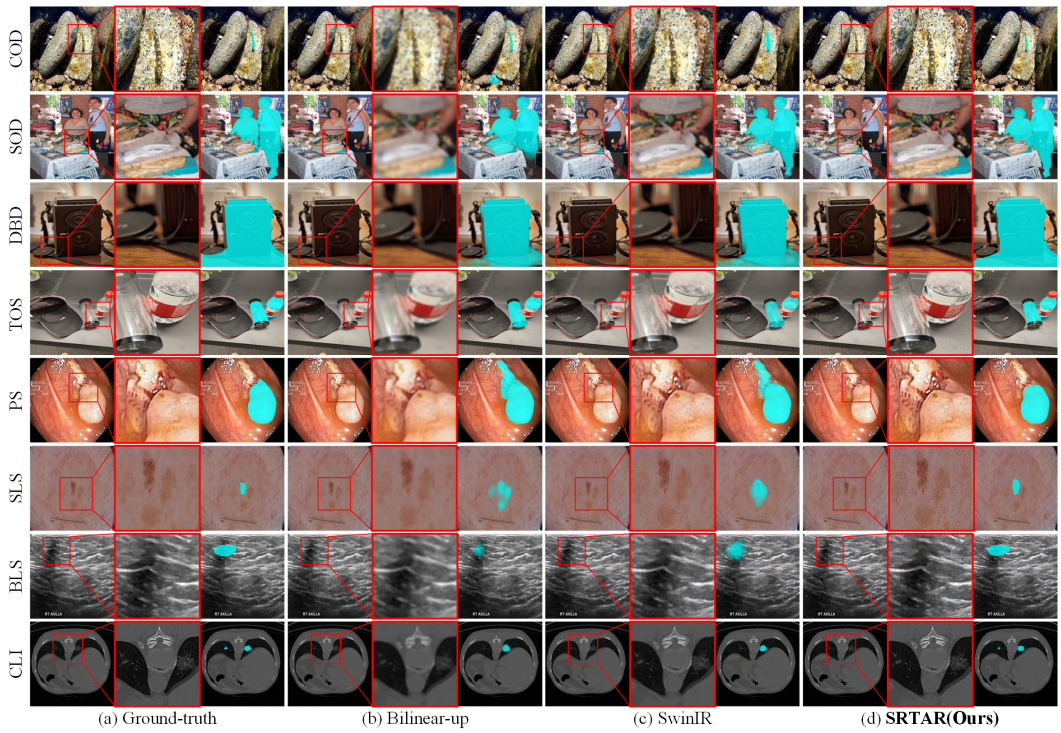


Figure 4: Qualitative comparison across eight challenging object segmentation tasks. We present the restored images and the corresponding predicted segmentation maps, respectively. Red boxes indicate enlarged local regions for detailed comparison. The SR scale factor is $\times 4$.

serve task-relevant structures. SRTAR provides more consistent gains across all tasks. Relative to $I_{\mathrm{S}} \rightarrow T$, it reduces Trans10K MAE by 42% and 35% under $\times 4$ and $\times 2$, respectively. Similar improvements occur on the remaining tasks, indicating that restoring high-frequency information relevant to the task is much more critical to improving task performance. The improvements are generally larger at $\times 4$, where LR inputs lose more structural and boundary information.

SRTAR$_{\mathrm{B}}$ remains close to SRTAR across all tasks. As shown in Table 2, SRTAR$_{\mathrm{B}}$ reduces FLOPs by 32.11% and 63.86% for 256^2 and 512^2 inputs, and achieves $2.58\times$ and $8.08\times$ inference speedups, respectively. This variant reduces inference latency and memory consumption by avoiding transformer-based SR computation while preserving most of the segmentation improvements achieved by SRTAR. Compared with task-specific methods, SRTAR and SRTAR$_{\mathrm{B}}$ achieve competitive performance across eight tasks, showing the general applicability of task-aware refinement.

Furthermore, as shown in Table 3, the proposed SRTAR achieves the best PSNR and LPIPS scores on a scale of $\times 4$ SR in SOD task. These results demonstrate that our SR results are not only beneficial from a task perspective but are also capable of producing visually appealing results.

4.4 Qualitative Results

Figure 4 compares SRTAR with $I_{\mathrm{B}} \rightarrow T$ and $I_{\mathrm{S}} \rightarrow T$ in eight binary segmentation tasks. SRTAR produces more accurate prediction maps with more complete object structures and sharper boundaries, showing that task-aware SR recovers discriminative details that benefit segmentation. Its reconstructed images also provide the most visually pleasing results among the compared settings.

4.5 Ablation Studies

We conduct ablations on COD using identical 256×256 inputs and 1024×1024 targets to evaluate each design and its internal configuration.

Effect of the SR training strategy. Table 4 examines SR adaptation from fully frozen D_0 to fully optimized D_6, with detailed definitions in Appendix A. Performance does not improve monotonically as more SR stages are adapted. D_2, which updates only `layers.3` and `conv_after_body`, achieves the best F_β^ω on CAMO and remains nearly optimal on COD10K. Selectively adapting the final residual group therefore transfers task-specific supervision into SR while largely preserving pretrained restoration capability, providing a favorable accuracy and adaptation balance. We adopt D_2 as the default strategy.

Effect of the TAR. In Table 5, the baseline adopts an FPN style network with LR inputs, while introducing the SR network yields a clear improvement, confirming the importance of restoring spatial details for segmentation. Building TAR without gradient guidance brings limited gains, indicating that directly injecting restored details is insufficient to make them reliably useful for the task. SR-gradient guidance consistently outperforms HR gradients alone, suggesting that gradients extracted from reconstructed SR provide a more consistent attention signal because SR is the only image source available at inference. Since SR gradients may contain inaccurate responses, full TAR uses HR gradients as a training-time reference to guide SR gradients within task-relevant regions, achieving the best MAE and F_β^ω. TAR therefore transfers useful HR high-frequency knowledge to SR during training.

Effect of the MQPS. The MQPS group in Table 5 fixes TAR while varying the injected source, including SR, HR, and Mix, and the construction strategy, including Resize and Crop. Mixing HR and SR consistently outperforms either source alone, showing that heterogeneous quality levels reduce reliance on quality-specific high-frequency cues. Stage-matched cropping further outperforms global resizing by preserving local boundaries and fine textures that spatial downscaling may attenuate. Consequently, Mix+Crop achieves the best overall performance.

Effect of the TFL. The TFL group retains complete TAR and MQPS while evaluating three complementary constraints. Here, $\mathcal{L}_c$, $\mathcal{L}_s$, and $\mathcal{L}_g$ denote $\mathcal{L}_{\text{consist}}$, $\mathcal{L}_{\text{similar}}$, and $\mathcal{L}_{\text{gdt}}$, respectively. Starting from $\mathcal{L}_c$, adding $\mathcal{L}_s$ improves both benchmarks, showing that local agreement alone cannot preserve the spatial organization of task-relevant representations. Further adding $\mathcal{L}_g$ yields the strongest performance, with F_β^ω reaching 0.877 on CAMO. The progressive gains confirm the effectiveness of TFL through complementary multi-level supervision.

Table 2: Efficiency analysis for our methods

Method	Scale	FLOPs↓	FPS↑	Latency↓
SRTAR	×4	1874 G	2.38	419 ms
SRTAR_B	×4	**1272** G	**6.15**	**163** ms
SRTAR	×2	3519 G	0.77	1304 ms
SRTAR_B	×2	**1272** G	**6.20**	**161** ms

Table 3: SR performance on SOD task.

Method	Scale	DUTS-TE		PASCAL-S	
		PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
B	×4	32.65	0.147	29.37	0.224
S	×4	34.53	0.132	30.49	0.206
SRTAR	×4	**35.00**	**0.121**	**30.82**	**0.196**

Table 4: Ablation study of SR training strategy on COD task.

Dataset	Metrics	D_0	D_1	D_2 (Ours)	D_3	D_4	D_5	D_6
CAMO	$\mathcal{M}\downarrow$	0.036	**0.032**	**0.032**	0.033	**0.032**	**0.032**	0.034
	$F_\beta^w\uparrow$	0.860	0.874	**0.877**	0.868	0.867	0.875	0.870
COD10K	$\mathcal{M}\downarrow$	0.018	**0.017**	**0.017**	0.018	**0.017**	0.018	**0.017**
	$F_\beta^w\uparrow$	0.824	0.835	0.839	0.837	0.832	0.830	**0.840**

Table 5: Ablation study of Structure, TAR, MQPS, and TFL components in the proposed SRTAR framework on challenging COD tasks. G_{HR} and G_{SR} denote the HR and SR gradient branches, and Mix denotes mixed HR and SR patches.

Dataset	Metrics	Structure settings		TAR settings				MQPS settings						TFL settings		
		Base.	SR	w/o $G_{\text{HR}}, G_{\text{SR}}$	w/o G_{SR}	w/o G_{HR}	TAR (Full)	SR + Resize	HR + Resize	Mix + Resize	SR + Crop	HR + Crop	Mix + Crop	$\mathcal{L}_p$	$\mathcal{L}_p$+$\mathcal{L}_s$	$\mathcal{L}_p$+$\mathcal{L}_s$+$\mathcal{L}_g$
CAMO	$\mathcal{M}\downarrow$	0.048	**0.039**	0.038	0.038	0.037	**0.035**	0.036	0.037	0.036	0.036	0.036	**0.035**	0.034	0.034	**0.032**
	$F_\beta^w\uparrow$	0.802	**0.851**	0.846	0.847	0.856	**0.865**	0.856	0.850	0.860	0.859	0.859	**0.865**	0.866	0.870	**0.877**
COD10K	$\mathcal{M}\downarrow$	0.028	**0.021**	0.021	0.020	0.020	**0.019**	**0.019**	0.020	**0.019**	**0.019**	**0.019**	**0.019**	0.019	0.018	**0.017**
	$F_\beta^w\uparrow$	0.728	**0.792**	0.800	0.806	0.816	**0.819**	0.815	0.811	0.818	0.815	**0.819**	**0.819**	0.822	0.827	**0.839**

5 Conclusion

We presented SRTAR, a task-aware super-resolution injection framework for challenging object segmentation from low-resolution inputs. Rather than treating SR as an isolated preprocessing step, SRTAR couples reconstructed detail with the segmentation pathway through TAR, regularizes cross-quality evidence using MQPS, and aligns prediction and SR representations with TFL. Experiments on eight natural-scene and medical segmentation tasks show outstanding performance against task-specialized methods.

AI USE STATEMENT

Generative AI tools were used in this work to assist with language polishing, manuscript formatting, and the retrieval and discovery of potentially relevant literature. All AI-assisted suggestions, including language revisions and retrieved references, were independently reviewed and verified by the authors before being incorporated into the manuscript. The authors take full responsibility for the accuracy, originality, claims, citations, and final content of this work.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. ISSN 2352-3409.

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, 2012.

Joan Bruna, Pablo Sprechmann, and Yann LeCun. Super-resolution with deep convolutional sufficient statistics. In *International Conference on Learning Representations*, 2016.

Shuyao Chai, Xixuan Zhao, Jiaming Zhang, and Jiangming Kan. Defocus blur detection based on transformer and complementary residual learning. *Multimedia Tools and Applications*, 83: 53095–53118, 2024.

Shijie Chang, Xiaoqi Zhao, Lihe Zhang, and Tiancheng Wang. Unified medical lesion segmentation via self-referring indicator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10414–10424, 2025.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision Workshops*, 2018.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the isic. *arXiv preprint arXiv:1902.03368*, 2019.

Ziteng Cui, Yingying Zhu, Lin Gu, Guo-Jun Qi, Xiaoxiao Li, Renrui Zhang, Zenghui Zhang, and Tatsuya Harada. Exploring resolution and degradation clues as self-supervised signal for low quality object detection. In *European Conference on Computer Vision*, pp. 473–491. Springer, 2022.

Dengxin Dai, Yujian Wang, Yuhua Chen, and Luc Van Gool. Is image super-resolution helpful for other vision tasks? In *IEEE Winter Conference on Applications of Computer Vision*, pp. 1–9, 2016. doi: 10.1109/WACV.2016.7477613.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Songsong Duan, Xi Yang, Nannan Wang, and Xinbo Gao. Lightweight rgb-d salient object detection from a speed-accuracy tradeoff perspective. *TIP*, 34:2529–2543, 2025.

Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4548–4557, 2017.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, pp. 2777–2787, 2020a.

Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 263–273, 2020b.

Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 39(8):2626–2637, 2020c.

Ke Fan, Changan Wang, Yabiao Wang, Chengjie Wang, Ran Yi, and Lizhuang Ma. Rfenet: Towards reciprocal feature evolution for glass segmentation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 717–725, 2023.

Jun Gao, Qicheng Lao, Qingbo Kang, Paul Liu, Chenlin Du, Kang Li, and Le Zhang. Boosting your context by dual similarity checkup for in-context learning medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(1):310–319, 2025.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2015.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020.

Chao Hao, Zitong Yu, Xin Liu, Jun Xu, Huanjing Yue, and Jingyu Yang. A simple yet effective network based on vision transformer for camouflaged object and salient object detection. *IEEE Transactions on Image Processing*, 34:608–622, 2025.

Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Task-driven super resolution: Object detection in low-resolution images. In *Neural Information Processing*, pp. 387–395. Springer, 2021. doi: 10.1007/978-3-030-92307-5_45.

Chunming He, Rihan Zhang, Fengyang Xiao, Chenyu Fang, Longxiang Tang, Yulun Zhang, Linghe Kong, Deng-Ping Fan, Kai Li, and Sina Farsiu. Run: Reversible unfolding network for concealed object segmentation. In *International Conference on Machine Learning (ICML)*, 2025.

Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5557–5566, 2023.

Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. Kvasir-seg: A segmented polyp dataset. In *ACM Multimedia*, pp. 451–462, 2020.

Wei Ji, Shunyu Yu, Jiaming Wu, Kaidong Ma, Cheng Bian, Qi Bi, Jianan Li, Hong Liu, Lecheng Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12341–12351, 2021.

Wei Ji, Shunyu Yu, Jiaming Wu, Kaidong Ma, Cheng Bian, Qi Bi, Jianan Li, Hong Liu, Lecheng Cheng, and Yefeng Zheng. Multispectral video semantic segmentation: A benchmark dataset and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1094–1104, 2023.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

Boce Kang. Pmrnet: Physics-informed multi-scale refinement network for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15659–15668, 2026.

Jaeha Kim, Junghun Oh, and Kyoung Mu Lee. Beyond image super-resolution for image recognition with task-driven perceptual loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2651–2661, 2024.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Taehun Kim, Hyemin Lee, and Daijin Kim. Uacanet: Uncertainty augmented context attention for polyp segmentation. In *ACM Multimedia*, pp. 2167–2175, 2021.

Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *CVIU*, 184:45–56, 2019.

Jianan Li, Wei Ji, Shuhe Wang, Wenguan Wang, and Lecheng Cheng. Dvsod: Rgb-d video salient object detection. In *Advances in Neural Information Processing Systems*, pp. 8774–8787, 2023.

Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *CVPR*, pp. 280–287, 2014.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

Jiaying Lin, Yuen-Hei Yeung, Shuquan Ye, and Rynson W. H. Lau. Leveraging rgb-d data with cross-modal context mining for glass surface detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5254–5261, 2025. Issue 5.

Ding Liu, Bihan Wen, Xianming Liu, Zhangyang Wang, and Thomas S. Huang. When image denoising meets high-level vision tasks: A deep learning approach. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 842–848, 2018. doi: 10.24963/ijcai.2018/117.

Nian Liu, Ziyang Luo, Ni Zhang, and Junwei Han. Vst++: Efficient and stronger visual saliency transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7300–7316, 2024a.

Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for universal foreground segmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48 (2):1762–1777, 2026.

Yutong Liu, Haijiang Zhu, Mengting Liu, Huaiyuan Yu, Zihan Chen, and Jie Gao. Rolling-unet: Revitalizing mlp's ability to efficiently extract long-distance dependencies for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3819–3827, 2024b. Issue 4.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hung-Ju Wang, Tinghui Zhou, Zhiqiang Shen, Bingyi Kang, Evan Shelhamer, and Trevor Darrell. Exploring simple and transferable recognition-aware image processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3032–3046, 2023. doi: 10.1109/TPAMI.2022.3183243.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, and Junwei Han. Vscode: General visual salient and camouflaged object detection with 2d prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17169–17180, 2024.

Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2014.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2160–2170, 2022.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9205–9220, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8024–8035, 2019.

Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11769–11779, 2024.

Peng Ren, Cheng Jiang, Fuming Sun, and Tian Bai. Depth-assisted camouflaged object segmentation via frequency-domain fusion and high-order interaction. *IEEE Transactions on Multimedia*, 28: 6865–6876, 2026.

Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

Jianping Shi, Li Xu, and Jiaya Jia. Discriminative blur detection features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Ke Sun, Zhongxi Chen, Xianming Lin, Xiaoshuai Sun, Hong Liu, and Rongrong Ji. Conditional diffusion models for camouflaged and salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2833–2848, 2025.

Yanguang Sun, Shuo Wang, Chenxi Chen, and Tian-Zhu Xiang. Boundary-guided camouflaged object detection. *arXiv preprint arXiv:2207.00794*, 2022.

Yanguang Sun, Chunyan Xu, Jian Yang, Hanyu Xuan, and Lei Luo. Frequency-spatial entanglement learning for camouflaged object detection. In *European Conference on Computer Vision (ECCV)*, pp. 343–360, 2024.

Nima Tajbakhsh, Suryakanth R. Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging*, 35(2):630–644, 2015.

Fenghe Tang, Lingtao Wang, Chunping Ning, Min Xian, and Jianrui Ding. Cmu-net: A strong convmixer-based medical ultrasound image segmentation network. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2023.

Li Wang, Dong Li, Yousong Zhu, Lu Tian, and Yi Shan. Dual super-resolution learning for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3774–3783, 2020.

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 *CVPR*, pp. 136–145, 2017.

Wei Wang, Huiying Sun, and Xin Wang. Lssnet: A method for colon polyp segmentation based on local feature supplementation and shallow feature supplementation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 446–456. Springer, 2024.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

Xiandong Wang, Fengqin Yao, Guoqiang Zhong, Qing Cai, Shengke Wang, and James Tin-Yau Kwok. Gbnet: Gated boundary-aware network for camouflaged object detection. *IEEE Transactions on Image Processing*, 35:5297–5310, 2026.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1130–1140, 2023.

Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In *ECCV*, pp. 696–711, 2020.

Zijin Yin, Kongming Liang, Zhanyu Ma, and Jun Guo. Duplex contextual relation network for polyp segmentation. In *ISBI*, pp. 1–5, 2022.

Guanghui Yue, Shangjie Wu, Tianwei Zhou, Gang Li, Jie Du, Yu Luo, and Qiuping Jiang. Progressive region-to-boundary exploration network for camouflaged object detection. *IEEE Transactions on Multimedia*, 27:236–248, 2025.

Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, 2010.

Mingfeng Zha, Guoqing Wang, Tianyu Li, Wei Dong, Peng Wang, and Yang Yang. Seeing beyond illusion: Generalized and efficient mirror detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12331–12339, 2026. Issue 15.

Chenxi Zhang, Qing Zhang, Jiayun Wu, and Youwei Pang. Cgcod: Class-guided camouflaged object detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4369–4377, 2025.

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention*, pp. 14–24, 2021.

Wenda Zhao, Fan Zhao, Dong Wang, and Huchuan Lu. Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3080–3088, June 2018a.

Wenda Zhao, Xueqing Hou, You He, and Huchuan Lu. Defocus blur detection via boosting diversity of deep ensemble networks. *IEEE Transactions on Image Processing*, 30:5426–5438, 2021a.

Wenda Zhao, Cai Shang, and Huchuan Lu. Self-generated defocus blur detection via dual adversarial discriminators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6933–6942, 2021b.

Wenda Zhao, Fei Wei, Haipeng Wang, You He, and Huchuan Lu. Full-scene defocus blur detection with defbd+ via multi-level distillation learning. *IEEE Transactions on Multimedia*, 25:9228–9240, 2023.

Xiaoqi Zhao, Youwei Pang, Wei Ji, Baicheng Sheng, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Spider: A unified framework for context-dependent concept segmentation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60906–60926. PMLR, 2024.

Xiaoqi Zhao, Hongpeng Jia, Youwei Pang, Long Lv, Feng Tian, Lihe Zhang, Weibing Sun, and Huchuan Lu. M^2SNet: Multi-scale in multi-scale subtraction network for medical image segmentation. *Machine Intelligence Research*, 2026.

Xiaotong Zhao, Wei Li, Yifan Zhang, and Zhiyong Feng. Residual super-resolution single shot network for low-resolution object detection. *IEEE Access*, 6:47780–47793, 2018b. doi: 10.1109/ACCESS.2018.2867586.

Liguo Zhou, Guang Chen, Mingyue Feng, and Alois C. Knoll. Improving low-resolution image classification by super-resolution with enhancing high-frequency content. In *25th International Conference on Pattern Recognition*, pp. 1972–1978, 2020.

Tao Zhou, Yi Zhou, Kelei He, Chen Gong, Jian Yang, Huazhu Fu, and Dinggang Shen. Cross-level feature aggregation network for polyp segmentation. *Pattern Recognition*, 140:109555, 2023a.

Tao Zhou, Yiyi Zhou, Kelei He, Chen Gong, Jian Yang, Huazhu Fu, and Dinggang Shen. Cross-level feature aggregation network for polyp segmentation. *Pattern Recognition*, 140:109555, 2023b.

A Additional Ablation Studies

Specifically, the SR adaptation strategies evaluated in Table 4 range from the fully frozen setting D_0 to the fully optimized setting D_6. Detailed definitions of these strategies are summarized in Table 6.

Table 6: Different SR fine-tuning strategies for SwinIR.

SR Training Strategy	Trainable SwinIR Modules
D0	None
D1	`conv_after_body`
D2	`layers.3` + `conv_after_body`
D3	`layers.2-3` + `conv_after_body`
D4	`layers.1-3` + `conv_after_body`
D5	`layers.0-3` + `conv_after_body`
D6	Entire SwinIR

B Dataset Details

B.1 Datasets and Metrics

Datasets. We evaluate SRTAR across eight binary segmentation tasks under widely adopted task-specific protocols. For COD, we jointly use CAMO-TR (Le et al., 2019) and COD10K-TR (Fan et al., 2020a) for training and evaluate the model on CAMO-TE and COD10K-TE. For SOD, we train the model on DUTS-TR (Wang et al., 2017) and evaluate it on DUTS-TE and PASCAL-S (Li et al., 2014). For PS, we combine the training partitions of CVC-ClinicDB (Tajbakhsh et al., 2015) and Kvasir-SEG (Jha et al., 2020), and report the performance on the test partition of Kvasir-SEG. For CLI, BLS, and SLS, we follow the standard data partitions of the dataset (Fan et al., 2020c), BUSI (Al-Dhabyani et al., 2020), and ISIC2018 (Codella et al., 2019), respectively. For TOS, we use the official training and test partitions of Trans10K (Xie et al., 2020). For DBD, we follow a cross-dataset evaluation protocol, in which the model is trained on the CUHK dataset (Shi et al., 2014) and evaluated on the 500-image DUT benchmark (Zhao et al., 2018a). Following standard super-resolution protocols (Agustsson & Timofte, 2017; Bevilacqua et al., 2012; Zeyde et al., 2010), we resize each high-resolution image to 1024×1024 and generate the corresponding low-resolution

input through bicubic downsampling. The input resolution is 256×256 for $\times 4$ SR and 512×512 for $\times 2$ SR.

Metrics. We use the task-specific evaluation metrics reported in Table 1 to verify the superiority of our model. For COD, we report structure-measure Fan et al. (2017) (S_m), weighted F-measure Margolin et al. (2014) (F_β^ω), and mean absolute error ($\mathcal{M}$). For SOD, we report S_m and $\mathcal{M}$. For the four medical image segmentation tasks, we report mean foreground Dice coefficient (Dice) and mean foreground Intersection over Union (IoU) for PS, BLS, and SLS, while Dice and mean IoU over the foreground and background (mIoU) for CLI. For TOS, we report IoU and $\mathcal{M}$. For DBD, we use the conventional F-measure (F_β) and $\mathcal{M}$. A higher value indicates better performance for every metric except MAE, for which a lower value is preferred.