When Specialists Meet Generalists: Segmenter-Coordinated Asymmetric Learning for Label-Deficient Concealed Object Segmentation

Chunming He[1] , Dingming Zhang[1] , Longxiang Tang[2] , Ziyun Yang[3] , Fengyang Xiao[1,†] , and Sina Farsiu[1,†]
[1]Duke University, [2]Harvard University, [3]Apple.

† Corresponding author. Contact: chunming.he@duke.edu

Abstract. Existing label-deficient concealed object segmentation (LDCOS) methods either rely on consistency constraints within a mean-teacher framework or employ the Segment Anything Model (SAM) as a fixed pseudo-label generator. Interestingly, task-specific segmenters and foundation models exhibit *complementary* failure modes: segmenters lack generalization under scarce labels, while SAM suffers from domain gaps in concealed scenes. A natural solution is to employ co-training to mutually promote both models. However, we observe that standard homogeneous co-training cannot exploit this complementarity because architecturally identical networks tend to share similar error patterns, especially when targets are heavily concealed. To address this, we present **SCALER** (Segmenter-Coordinated Asymmetric LEaRning), a framework that jointly optimizes a mean-teacher segmenter and a learnable SAM through two alternating phases with *model-specific* optimization strategies. In **Phase I**, the segmenter is updated under fixed SAM supervision using entropy-based image-level and uncertainty-based pixel-level weighting to suppress unreliable pseudo-label regions. In **Phase II**, SAM is updated via an augmentation invariance loss and a noise resistance loss, which exploit SAM's inherent perturbation robustness rather than treating it identically to the segmenter. This asymmetric design is the key distinction from prior co-training and one-way distillation methods: each model's learning strategy is tailored to its own inductive bias, enabling mutual enhancement. Experiments across eight LDCOS tasks demonstrate consistent gains, and SCALER improves both the lightweight segmenter and the foundation model, serving as a general plug-and-play paradigm for label-scarce conditions. The code is available at https://github.com/ChunmingHe/SCALER.

Keywords: Label-deficient concealed object segmentation · Foundation models · Co-training framework

1 Introduction

Concealed object segmentation contains scenarios where targets share high visual similarity with their surroundings, making them difficult to distinguish. Repre-

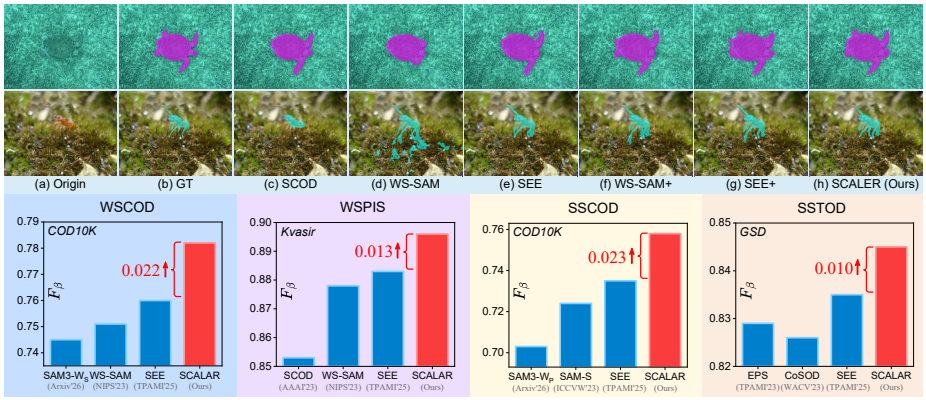

Fig. 1: Results of LDCOS methods, including SCOD [23], WS-SAM [16], and SEE [17]. The suffix "+" means integration with SCALER. SCALER achieves leading results across COD, PIS, and TOD under weak (WS) and semi-supervised (SS) settings. In the top section, concealed objects masks are highlighted in pink and blue.

sentative tasks include camouflaged object detection (COD) [12,22], polyp image segmentation (PIS) [21, 44], and transparent object detection (TOD) [43, 50].

The challenge intensifies in label-deficient concealed object segmentation (LDCOS), covering semi-supervised (SSCOS) or weakly-supervised (WSCOS) setups. Prior works have extended consistency-based frameworks, particularly the mean-teacher paradigm [23, 28, 41] that enforces prediction stability between a student and an exponentially averaged teacher. However, pseudo-labels (PLs) generated purely by the mean-teacher are prone to error accumulation, especially for concealed targets with uncertain boundaries (see Fig. 1).

To incorporate foundation model guidance for more reliable PLs, recent works such as WS-SAM [16] and SEE [17] employed the pretrained Segment Anything Model (SAM) [27] as an external pseudo-label source. Despite SAM's value in supplementary supervision, they face two limitations (see Figs. 2 and 3): (1) SAM's fixed predictions are not adapted for COS, yielding noisy pseudo-labels. (2) The information flow is strictly one-way, from SAM to the segmenter, with no mechanism for the domain-adapted segmenter to reciprocally improve SAM.

A natural idea is to let both models teach each other, analogous to co-training [8,14,19,36,42]. However, conventional co-training methods use architecturally identical networks, which tend to share similar error patterns and suffer from confirmation bias. In COS, where concealed targets are inherently ambiguous, such homogeneous pairing amplifies these issues. In contrast, a task-specific segmenter and a foundation model, like SAM series [27, 32, 34], possess fundamentally different inductive biases: the segmenter is sensitive to label noise but excels at domain adaptation, while SAM is robust to perturbations but suffers from domain gaps. This complementarity suggests that an *asymmetric* collaborative strategy, where each model's learning procedure is tailored to its own properties, can be more effective than symmetric co-training.

Motivated by this observation, we propose **SCALER**, a Segmenter-Coordinated Asymmetric LEaRning framework for LDCOS. SCALER integrates a mean-

Fig. 2: Pseudo-labels generated by different SAM variants, including fixed SAM (SAM), fine-tuned SAM with SAMAdapter [7] (SAM+A), and SAM trained under our SCALER (SAM+S). "SAM+A" and "SAM+S" are trained on scribble supervision.

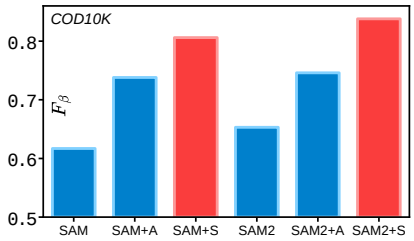

Fig. 3: Quality of pseudo-labels generated by different SAM variants, with the same definitions in Fig. 2.

teacher segmenter with a learnable SAM, optimized through two alternating phases with *model-specific* strategies.

In **Phase I**, we train the segmenter within the mean-teacher framework while keeping SAM fixed. Given the segmenter's sensitivity to label noise, we introduce entropy-based image-level weighting and uncertainty-based pixel-level weighting to selectively leverage pseudo-labels from both the teacher network and SAM. For samples already well-segmented, we apply strong augmentations to mine potentially informative features that may otherwise be overlooked.

In **Phase II**, we fix the segmenter and use it as a domain specialist to generate pseudo-labels for updating SAM. Critically, we do *not* reuse the segmenter's learning strategy. Instead, we design two losses that exploit SAM's distinct properties: an augmentation invariance loss that leverages SAM's pre-trained robustness to geometric perturbations, and a noise resistance loss that applies only pixel-level filtering so SAM can learn from all samples including difficult ones. This asymmetric treatment is the core distinction from both one-way distillation [16, 17] and standard co-training [8, 14].

By alternately optimizing both phases, SCALER progressively improves both models through mutual knowledge transfer. Our contributions are:

(1) We propose SCALER, an asymmetric collaborative learning framework that enables a task-specific segmenter and a foundation model (such as SAM) to mutually enhance each other in LDCOS. Unlike homogeneous co-training, SCALER assigns model-specific optimization strategies based on each model's inductive bias, yielding gains beyond architectural complementarity alone.

(2) We design two specialized optimization phases. For the segmenter (Phase I), we introduce entropy- and uncertainty-based weighting to filter pseudo-label noise. For SAM (Phase II), we propose augmentation invariance and noise resistance losses that exploit SAM's perturbation robustness, enabling it to learn effectively from noisy specialist feedback without sacrificing generalization.

(3) Extensive experiments on eight tasks validate SCALER's effectiveness and generalization. SCALER simultaneously enhances both lightweight segmenters and foundation models (SAM, SAM2, SAM3), and serves as a plug-and-play framework compatible with various segmenters and SAM adaptation methods.

2 Related Work

Concealed Object Segmentation. Learning-based segmenters have achieved success in COS under full supervision [12, 15, 18, 20, 47]. Representative methods include SINet [12] for reversible segmentation, RUN [20] for model-driven optimization, SCLoss [47] for spatial connectivity. Recent LDCOS techniques, including semi- and weakly-supervised setups, mainly employ consistency regularization [23, 28] or SAM-based pseudo-labeling [16, 17, 24]. However, unstable consistency and task-agnostic SAM supervision often lead to suboptimal results. SCALER unifies consistency regularization with adaptive SAM-based learning through an asymmetric collaborative design that achieves mutual refinement.

Label-Deficient Image Segmentation. Semi- and weakly-supervised segmentation improve label efficiency through regularization and pseudo-label refinement [1,5,29,35]. The mean-teacher framework [37], where a teacher updated via exponential moving average provides temporally ensembled targets, is a widely adopted paradigm. However, in COS, ambiguous boundaries and low contrast destabilize pseudo-labels and amplify error accumulation. Meanwhile, fixed or lightly tuned foundation models such as SAM [27] suffer domain gaps in concealed scenes. Recent methods WS-SAM [16] and SEE [17] design pseudo-label filtering strategies for SAM outputs, but remain restricted by SAM's limited capacity in concealed domains and one-way information flow.

Collaborative and Co-training Methods. Co-training leverages multiple models to generate pseudo-labels for each other. Representative methods include Co-teaching [14], CPS [8], and UCMT [36], which progressively extend mutual learning from small-loss sample selection to consistency enforcement and uncertainty-guided mixing. In COS, MiNet [33] and MIR-Net [48] explore bidirectional interaction between region and edge cues for weakly-supervised detection. However, these methods predominantly use *homogeneous* architectures, leading to correlated failure modes on concealed targets. SCALER departs from this paradigm by pairing a task-specific segmenter with a foundation model in a *heterogeneous* architecture, and crucially employing *asymmetric* optimization with model-specific loss functions tailored to each model's inductive bias.

3 Methodology

Label-deficient concealed object segmentation (LDCOS), including semi-supervised (SSCOS) and weakly supervised (WSCOS) setups, aims to train a segmenter from a partially labeled set $\mathcal{S} = \{X_i, Y_i\}_{i=1}^{S}$ and evaluate on a test set $\mathcal{T} = \{T_i\}_{i=1}^{T}$, where X_i and T_i denote input images and Y_i denotes the corresponding deficient label. In SSCOS, Y_i is missing for some training samples, whereas in WSCOS, Y_i comprises sparse annotations (*e.g.*, points or scribbles).

Training LDCOS models is challenging: consistency constraints alone are unstable for concealed targets, and SAM [27] provides general priors but suffers from domain gaps. Prior methods [16, 17] rely on *one-way* transfer from SAM to the segmenter, lacking a reciprocal mechanism. A straightforward alternative is homogeneous co-training [8, 14], but identical networks share correlated errors

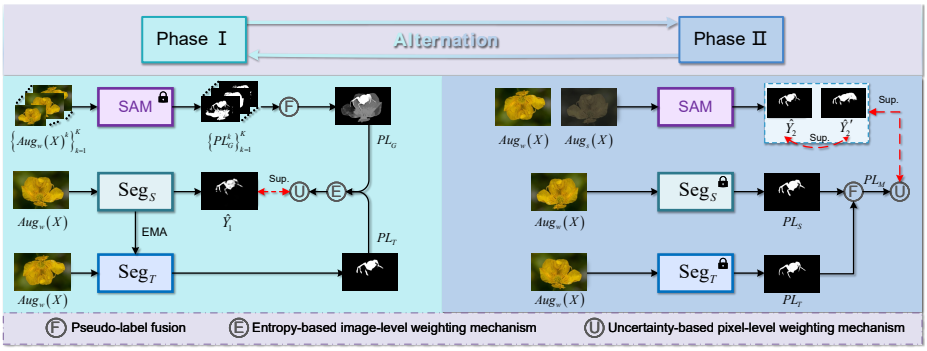

Fig. 4: The SCALER framework: asymmetric collaborative learning between a specialist segmenter and a generalist foundation model. The framework alternates between two optimization phases at each training batch. The lock icon indicates frozen parameters. Notably, our framework supports the integration of various existing segmenters and SAM/SAM2/SAM3 fine-tuning approaches.

on concealed objects. In contrast, a task-specific segmenter and SAM possess complementary failure modes: the segmenter adapts well but is noise-sensitive, while SAM generalizes broadly but lacks domain specificity. This complementarity motivates our *asymmetric* design.

We propose **SCALER**, a Segmenter-Coordinated Asymmetric LEaRning framework. As shown in Fig. 4, SCALER integrates a mean-teacher framework with a learnable SAM through alternating optimization: the segmenter learns from SAM (Phase I) while SAM learns from the segmenter (Phase II), with each phase using distinct loss functions matched to the target model's inductive bias.

Notation convention. Throughout this paper, we use subscript G for the generalist (SAM) model outputs, subscript S for the student segmenter, subscript T for the teacher segmenter, and superscript (i) for pixel indices. PL denotes pseudo-labels and $\hat{Y}$ denotes predictions.

3.1 Phase I: Training Segmenter with Fixed SAM

In Phase I, we fix SAM (as a "generalist" teacher) and optimize the segmenter. The segmenter benefits from dual-source supervision: (1) consistency regularization from its own teacher and (2) general knowledge from the frozen SAM.

Dual pseudo-label generation. We generate two pseudo-labels. First, for the teacher model Seg_T, we apply a random weak augmentation $Aug_w(\cdot)$ to each input X and obtain the teacher pseudo-label from the teacher model (Seg_T):

$$PL_T = \text{Seg}_T\left(Aug_w(X)\right). \tag{1}$$

Next, for SAM, we generate K stochastic augmentations $\{Aug_w(X)^k\}_{k=1}^K$ by randomly sampling flips, rotations (0°, 90°, 180°, 270°), and scales (×0.5, ×1.0, ×2.0). Feeding these variants into SAM yields:

$$\{PL_G^k\}_{k=1}^K = \text{SAM}(\{Aug_w(X)^k\}_{k=1}^K), \tag{2}$$

Algorithm 1 SCALER: Iterative Mutual Enhancement (Semi-Supervised)

Require: Pre-trained $\text{Seg}_S(\theta_S)$, $\text{Seg}_T(\theta_T)$, SAM(ϕ); labeled $\mathcal{D}_L$, unlabeled $\mathcal{D}_U$; EMA decay η.

for each training iteration **do**
  Sample $(X_l, Y_l) \sim \mathcal{D}_L$, $X_u \sim \mathcal{D}_U$; $X \leftarrow \text{Concat}(X_l, X_u)$
  // Phase I: Update Segmenter (SAM fixed)
  $PL_T \leftarrow \text{Seg}_T(Aug_w(X))$; $PL_G \leftarrow \text{Ensemble-SAM}(X)$
  $\hat{Y}_S \leftarrow \text{Seg}_S(Aug_w(X))$
  $\mathcal{L}_s^1 \leftarrow R(PL_T, \hat{Y}_S) + R(PL_G, \hat{Y}_S) + \mathcal{L}_{ce}(\hat{Y}_S, Y_l) + \mathcal{L}_{IoU}(\hat{Y}_S, Y_l)$
  $\theta_S \leftarrow \theta_S - \alpha \nabla_{\theta_S} \mathcal{L}_s^1$; $\theta_T \leftarrow \eta\theta_T + (1-\eta)\theta_S$
  // Phase II: Update SAM (Segmenters fixed)
  $PL_M \leftarrow (\text{Seg}_S(Aug_w(X)) + \text{Seg}_T(Aug_w(X)))/2$
  $\hat{Y}_G \leftarrow \text{SAM}(Aug_w(X))$; $\hat{Y}'_G \leftarrow \text{SAM}(Aug_s(X))$
  $\mathcal{L}_s^2 \leftarrow \mathcal{L}_{ai}(\hat{Y}_G, \hat{Y}'_G) + \mathcal{L}_{nr}(\hat{Y}_G, PL_M) + \mathcal{L}_{ce}(\hat{Y}_G, Y_l) + \mathcal{L}_{IoU}(\hat{Y}_G, Y_l)$
  $\phi \leftarrow \phi - \beta \nabla_\phi \mathcal{L}_s^2$
end for

where we use the subscript G for SAM outputs. We aggregate by averaging:

$$PL_G = \frac{1}{K} \sum_{k=1}^{K} PL_G^k. \tag{3}$$

This ensemble leverages SAM's sensitivity to subtle variations [17], yielding a PL_G that is generally more accurate than any single prediction.

Pseudo-label refinement. Initial pseudo-labels may be unreliable at both the sample and pixel levels. We propose two complementary weighting mechanisms: an entropy-based image-level weight and an uncertainty-based pixel-level weight. The image-level entropy for a pseudo-label map PL is:

$$E(PL) = \frac{1}{N} \sum_{i=1}^{N} -PL^{(i)} \log_2 PL^{(i)} - (1 - PL^{(i)}) \log_2 (1 - PL^{(i)}), \tag{4}$$

where N is the number of pixels and superscript (i) indexes pixels. High entropy indicates low overall sample confidence. The pixel-level uncertainty weight is:

$$U^{(i)}(PL) = (2\, PL^{(i)} - 1)^2, \tag{5}$$

which attenuates pixels with predictions near 0.5, reflecting ambiguity in foreground-background discrimination. Both are computed for PL_T and PL_G.

Unlike prior methods [16,17] that simply combine these weights, we define a spatially-weighted basic loss $R_b(\bullet)$ that properly integrates pixel-level uncertainty *before* aggregation. Given a pseudo-label PL, a prediction $\hat{Y}$, and a spatial weight map W (defaulting to $U(PL)$), we define:

$$R_b(PL, \hat{Y}, W) = (1 - E(PL)) \cdot \Big(\mathcal{L}_{ce}^w(\hat{Y}, PL, W) + \mathcal{L}_{IoU}^w(\hat{Y}, PL, W) \Big), \tag{6}$$

where $\mathcal{L}_{ce}^w$ and $\mathcal{L}_{IoU}^w$ are the spatially-weighted cross-entropy and IoU losses:

$$\mathcal{L}_{ce}^w(\hat{Y}, PL, W) = -\frac{\sum_i W^{(i)} \left[PL^{(i)} \log \hat{Y}^{(i)} + (1 - PL^{(i)}) \log(1 - \hat{Y}^{(i)}) \right]}{\sum_i W^{(i)}}, \tag{7}$$

$$\mathcal{L}_{IoU}^w(\hat{Y}, PL, W) = 1 - \frac{\sum_i W^{(i)} \cdot \hat{Y}^{(i)} \cdot PL^{(i)}}{\sum_i W^{(i)} \cdot (\hat{Y}^{(i)} + PL^{(i)} - \hat{Y}^{(i)} \cdot PL^{(i)})}. \tag{8}$$

When called as $R_b(PL, \hat{Y})$ without an explicit weight argument, the default weight $W = U(PL)$ is used. Here $(1 - E(PL))$ provides sample-level confidence scaling, while W provides pixel-level weighting integrated *inside* the loss.

The student segmenter prediction is $\hat{Y}_S = \text{Seg}_S(Aug_w(X))$. Rather than treating all samples uniformly, we further address two extreme cases:
(1) Difficult samples: If $E(PL_T) \geq 0.8$, indicating low overall confidence, we introduce a binary mask $\mathbf{m}_{PL_T}$ to retain only highly certain pixels:

$$\mathbf{m}_{PL_T}^{(i)} = \begin{cases} 0 & \text{if } PL_T^{(i)} \in [0.1, 0.9], \\ 1 & \text{otherwise.} \end{cases} \tag{9}$$

The refined loss is $R(PL_T, \hat{Y}_S) = R_b(PL_T, \hat{Y}_S, \mathbf{m}_{PL_T} \odot U(PL_T))$, where $\odot$ denotes the Hadamard product, combining the binary mask with pixel-level uncertainty.
(2) Simple samples: If $E(\hat{Y}_S) \leq 0.2$, we apply strong augmentations $Aug_s(\cdot)$ to mine potentially informative features, obtaining $\hat{Y}'_S = \text{Seg}_S(Aug_s(X))$:

$$R(PL_T, \hat{Y}_S) = R_b(PL_T, \hat{Y}_S) + R_b(PL_T, \hat{Y}'_S). \tag{10}$$

Overall, the piecewise refinement strategy is:

$$R(PL_T, \hat{Y}_S) = \begin{cases} R_b(PL_T, \hat{Y}_S,\ \mathbf{m}_{PL_T} \odot U(PL_T)), & E(PL_T) \geq 0.8, \\ R_b(PL_T, \hat{Y}_S) + R_b(PL_T, \hat{Y}'_S), & E(\hat{Y}_S) \leq 0.2, \\ R_b(PL_T, \hat{Y}_S), & \text{otherwise.} \end{cases} \tag{11}$$

Optimization. Under weak supervision, the segmenter is trained with:

$$\mathcal{L}_w^1 = R(PL_T, \hat{Y}_S) + R(PL_G, \hat{Y}_S) + \mathcal{L}_{pce}(Y, \hat{Y}_S), \tag{12}$$

where $\mathcal{L}_{pce}$ is the partial cross-entropy loss. Under semi-supervision, $\mathcal{L}_{pce}$ is replaced by $\mathcal{L}_{ce} + \mathcal{L}_{IoU}$ for labeled samples, while unlabeled samples use the refinement loss, formulated as:

$$\mathcal{L}_s^1 = R(PL_T, \hat{Y}_S) + R(PL_G, \hat{Y}_S) + \mathcal{L}_{ce}(Y, \hat{Y}_S) + \mathcal{L}_{IoU}(Y, \hat{Y}_S). \tag{13}$$

The teacher weights θ_T are updated via exponential moving average (EMA):

$$\theta_T \leftarrow \eta\theta_T + (1 - \eta)\theta_S, \quad \text{with } \eta = 0.996. \tag{14}$$

3.2 Phase II: Optimizing SAM via Segmenter Feedback

This phase enables the mutual learning loop. We reverse the roles: the specialized mean-teacher framework $(\text{Seg}_S, \text{Seg}_T)$ is now *fixed* and serves as a "specialist" teacher to update the learnable SAM (via SAMadapter [7]).
Why asymmetric optimization? A naive approach would reuse Phase I's strategy for SAM. However, SAM and the segmenter have fundamentally different properties. SAM's prompt-based, scale-agnostic pretraining makes it inherently robust to geometric and photometric perturbations, which the segmenter lacks. Meanwhile, the segmenter excels at domain-specific discrimination but is sensitive to noise. Therefore, Phase II's losses are designed to *exploit* SAM's perturbation robustness rather than to protect against noise as in Phase I. This asymmetry is what distinguishes SCALER from standard co-training, where both peers receive identical treatment. Importantly, the principle of tailoring optimization to each model's inductive bias is not specific to SAM; it naturally extends to any foundation model with similar pretrained robustness.

We generate a consensus pseudo-label from the fixed segmenter: $PL_M = (PL_S + PL_T)/2$, where $PL_S = \text{Seg}_S(Aug_w(X))$ and $PL_T = \text{Seg}_T(Aug_w(X))$. We then design two losses for robust knowledge transfer.

Augmentation Invariance Loss ($\mathcal{L}_{ai}$). SAM's pretraining endows it with strong invariance to spatial transformations. We reinforce this property by enforcing prediction consistency between weak and strong augmentations:

$$\mathcal{L}_{ai} = \mathcal{L}_{ce}(\hat{Y}_G, \hat{Y}'_G) + \mathcal{L}_{IoU}(\hat{Y}_G, \hat{Y}'_G), \tag{15}$$

where $\hat{Y}_G = \text{SAM}(Aug_w(X))$ and $\hat{Y}'_G = \text{SAM}(Aug_s(X))$. By leveraging SAM's existing invariance as a regularizer rather than imposing it from scratch, this loss facilitates robust adaptation without impairing generalization.

Noise Resistance Loss ($\mathcal{L}_{nr}$). This loss distills knowledge from PL_M while protecting SAM from its noise. Unlike Phase I, we apply only pixel-level uncertainty weighting $U(\cdot)$ to ignore ambiguous pixels, deliberately omitting the image-level entropy weight $E(\cdot)$. This ensures SAM learns from *all* samples, including difficult ones that would destabilize the segmenter:

$$\mathcal{L}_{nr} = \mathcal{L}^w_{ce}(\hat{Y}_G, PL_M, U(PL_M)) + \mathcal{L}^w_{IoU}(\hat{Y}_G, PL_M, U(PL_M)), \tag{16}$$

where $\mathcal{L}^w_{ce}$ and $\mathcal{L}^w_{IoU}$ are defined in Eqs. (7) and (8). The two losses jointly increase training diversity and mitigate overfitting under limited supervision.

Optimization. Under weak supervision, SAM is updated via:

$$\mathcal{L}^2_w = \mathcal{L}_{ai} + \mathcal{L}_{nr} + \mathcal{L}_{pce}(Y, \hat{Y}_G). \tag{17}$$

Under semi-supervision:

$$\mathcal{L}^2_s = \mathcal{L}_{ai} + \mathcal{L}_{nr} + \mathcal{L}_{ce}(Y, \hat{Y}_G) + \mathcal{L}_{IoU}(Y, \hat{Y}_G). \tag{18}$$

3.3 Optimization Procedure

The training loop alternates Phase I and Phase II *within each batch*, enabling tight coupling so each model immediately benefits from the other's latest updates. At inference, only the student segmenter Seg_S is used, incurring no additional cost. The complete procedure is summarized in Algorithm 1.

3.4 Convergence Analysis

SCALER's batch-level Phase I/Phase II alternation can be cast as block coordinate descent [38, 45] on the joint energy $E(\theta, \phi) = \mathcal{L}_{\text{I}}(\theta; \phi) + \lambda \mathcal{L}_{\text{II}}(\phi; \theta)$, where θ and ϕ denote the segmenter and SAM parameters. Under standard assumptions (L-smoothness of each loss in its optimized variable, boundedness of E from below, and step sizes $\alpha \leq 1/L_1$, $\beta \leq \lambda/L_2$), we establish:

Proposition 1 (Convergence of SCALER). *The iterates $\{(\theta^t, \phi^t)\}$ satisfy: (i) E decreases up to a bounded perturbation $\delta^t \to 0$, with per-step descent at least $\frac{c}{2}(\|\Delta\theta^t\|^2 + \|\Delta\phi^t\|^2) - \delta^t$; (ii) $\min_{t \leq T} \|\nabla E^t\|^2 = \mathcal{O}(1/T)$, so the iterates converge to an ε-approximate stationary point with $\varepsilon \to 0$.*

The proof follows from bounding each phase's descent via L-smoothness, controlling the cross-term coupling, and telescoping. The piecewise refinement is formulated as a continuous sigmoid relaxation, preserving smoothness. Two sources of

Table 1: Results on COD of the WSCOS task with different supervision, segmenter backbones, and foundation models. The best two results are in **red** and **blue** fonts.

Methods	Pub.	*CHAMELEON*				*CAMO*				*COD10K*				*NC4K*			
		$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
Scribble Supervision, ResNet50, SAM																	
SAM [27]	ICCV23	0.207	0.595	0.647	0.635	0.160	0.597	0.639	0.643	0.093	0.673	0.737	0.730	0.118	0.675	0.723	0.717
SAM-W$_S$ [7]	ICCVW23	0.069	0.751	0.835	0.661	0.097	0.696	0.788	0.738	0.049	0.712	0.833	0.770	0.066	0.757	0.842	0.768
TEL [30]	CVPR22	0.073	0.708	0.827	0.785	0.104	0.681	0.797	0.717	0.057	0.633	0.826	0.724	0.075	0.754	0.832	0.782
SCOD [23]	AAAI23	0.046	**0.791**	0.897	0.818	0.092	0.709	0.815	0.735	0.049	0.637	0.832	0.733	0.064	0.751	0.853	0.779
WS-SAM [16]	NIPS23	0.046	0.777	0.897	0.824	0.092	0.742	0.818	0.759	0.038	0.719	0.878	0.803	0.052	0.802	0.886	0.829
SEE [17]	TPAMI25	**0.044**	0.785	**0.903**	**0.826**	**0.090**	**0.747**	**0.826**	**0.765**	**0.036**	**0.729**	**0.883**	**0.807**	**0.051**	**0.808**	**0.891**	**0.836**
SCALER (Ours)	—	**0.042**	**0.792**	**0.915**	**0.838**	**0.089**	**0.784**	**0.835**	**0.777**	**0.034**	**0.737**	**0.885**	**0.814**	**0.050**	**0.811**	**0.898**	**0.837**
Scribble Supervision, PVT V2, SAM3																	
SAM3-W$_S$ [6]	Arxiv 26	0.067	0.766	0.839	0.667	0.095	0.703	0.790	0.742	0.047	0.745	0.866	0.772	0.063	0.783	0.856	0.775
WS-SAM [16]	NIPS23	0.041	0.787	0.902	0.838	0.080	0.779	0.850	0.793	0.032	0.751	0.893	0.827	0.045	0.825	0.901	0.842
SEE [17]	TPAMI25	**0.039**	**0.815**	**0.919**	**0.840**	**0.076**	**0.803**	**0.858**	**0.800**	**0.030**	**0.760**	**0.902**	**0.832**	**0.045**	**0.831**	**0.913**	**0.837**
SCALER (Ours)	—	**0.032**	**0.830**	**0.941**	**0.852**	**0.071**	**0.828**	**0.875**	**0.819**	**0.026**	**0.782**	**0.917**	**0.842**	**0.042**	**0.843**	**0.927**	**0.853**
Point Supervision, ResNet50, SAM																	
SAM [27]	ICCV23	0.207	0.595	0.647	0.635	0.160	0.597	0.639	0.643	0.093	0.673	0.737	0.730	0.118	0.675	0.723	0.717
SAM-W$_P$ [7]	ICCVW23	0.095	0.708	0.752	0.688	0.117	0.653	0.698	0.681	0.068	0.697	0.805	0.767	0.078	0.736	0.793	0.782
TEL [30]	CVPR22	0.094	0.712	0.751	0.746	0.133	0.662	0.674	0.645	0.063	0.623	0.803	0.727	0.085	0.725	0.795	0.766
SCOD [23]	AAAI23	0.092	0.688	0.746	0.725	0.137	0.629	0.688	0.663	0.060	0.607	0.802	0.711	0.080	0.744	0.796	0.758
WS-SAM [16]	NIPS23	0.056	0.767	0.868	0.805	0.102	0.703	0.757	0.718	0.039	0.698	0.856	0.790	0.057	0.801	0.859	0.813
SEE [17]	TPAMI25	**0.055**	**0.772**	**0.872**	**0.806**	**0.098**	**0.712**	**0.769**	**0.721**	**0.038**	**0.706**	**0.862**	**0.796**	**0.055**	**0.806**	**0.867**	**0.817**
SCALER (Ours)	—	**0.054**	**0.781**	**0.872**	**0.810**	**0.096**	**0.718**	**0.779**	**0.730**	**0.038**	**0.719**	**0.876**	**0.801**	**0.053**	**0.812**	**0.869**	**0.822**
Point Supervision, PVT V2, SAM3																	
SAM3-W$_P$ [6]	Arxiv 26	0.093	0.715	0.754	0.690	0.112	0.656	0.701	0.684	0.066	0.703	0.808	0.771	0.076	0.742	0.795	0.784
WS-SAM [16]	NIPS23	0.051	0.792	0.880	0.822	0.089	0.743	0.788	0.728	0.035	0.724	0.865	0.808	0.050	0.822	0.870	0.826
SEE [17]	TPAMI25	**0.043**	**0.795**	**0.883**	**0.824**	**0.085**	**0.750**	**0.800**	**0.734**	**0.033**	**0.735**	**0.866**	**0.814**	**0.048**	**0.833**	**0.883**	**0.836**
SCALER (Ours)	—	**0.035**	**0.826**	**0.903**	**0.840**	**0.070**	**0.791**	**0.831**	**0.753**	**0.029**	**0.758**	**0.897**	**0.827**	**0.044**	**0.859**	**0.898**	**0.849**

Fig. 5: Visualizations for COS tasks with point supervision.

non-stationarity, the stop-gradient pseudo-labels and the EMA teacher θ_T, are bounded as vanishing perturbations via Bertsekas [2]. The $\mathcal{O}(1/T)$ rate is empirically consistent with Fig. 6. Global convergence follows from the KŁ property [4], which is satisfied for losses definable in an o-minimal structure [3].

Table 2: Results on COD of the SSCOS task with 1/8 and 1/16 labeled training data.

Methods	Pub.	*CHAMELEON*				*CAMO*				*COD10K*				*NC4K*			
		$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
1/8 Labeled Training Data, ResNet50, SAM																	
SAM [27]	ICCV23	0.207	0.595	0.647	0.635	0.160	0.597	0.639	0.643	0.093	0.673	0.737	0.730	0.118	0.675	0.723	0.717
SAM-S [7]	ICCVW23	0.136	0.667	0.695	0.672	0.133	0.637	0.678	0.662	0.073	0.695	0.770	0.751	0.085	0.715	0.756	0.767
PGCL [1]	CVPR23	0.051	0.792	0.878	0.833	0.096	0.705	0.803	0.755	0.051	0.658	0.798	0.752	0.063	0.753	0.838	0.803
EPS [29]	TPAMI23	0.042	0.810	0.891	0.862	0.090	0.721	0.815	0.753	0.047	0.659	0.806	0.761	0.058	0.765	0.855	0.809
CoSOD [5]	WACV24	0.047	0.802	0.883	0.850	0.092	0.730	0.822	0.761	0.046	0.673	0.813	0.767	0.061	0.764	0.862	0.818
SEE [17]	TPAMI25	0.035	0.825	0.903	0.873	0.083	0.753	0.843	0.776	0.040	0.703	0.839	0.786	0.053	0.778	0.889	0.839
SCALER (Ours)	—	**0.033**	**0.838**	**0.912**	**0.877**	**0.081**	**0.759**	**0.844**	**0.779**	**0.039**	**0.716**	**0.846**	**0.787**	**0.052**	**0.784**	**0.891**	**0.848**
1/8 Labeled Training Data, PVT V2, SAM3																	
SAM3-S [6]	Arxiv 26	0.133	0.672	0.686	0.675	0.131	0.645	0.687	0.669	0.071	0.702	0.775	0.757	0.083	0.723	0.760	0.767
CoSOD [5]	WACV24	0.044	0.817	0.892	0.860	0.080	0.752	0.835	0.774	0.044	0.692	0.825	0.783	0.057	0.783	0.882	0.826
SEE [17]	TPAMI25	0.033	0.840	0.911	0.881	0.080	0.765	0.856	0.789	0.037	0.727	0.851	0.802	0.049	0.797	0.902	0.848
SCALER (Ours)	—	**0.028**	**0.862**	**0.925**	**0.891**	**0.068**	**0.789**	**0.868**	**0.802**	**0.030**	**0.753**	**0.869**	**0.820**	**0.042**	**0.823**	**0.908**	**0.863**
1/16 Labeled Training Data, ResNet50, SAM																	
SAM [27]	ICCV23	0.207	0.595	0.647	0.635	0.160	0.597	0.639	0.643	0.093	0.673	0.737	0.730	0.118	0.675	0.723	0.717
SAM-S [7]	ICCVW23	0.150	0.642	0.680	0.657	0.146	0.624	0.663	0.647	0.078	0.682	0.752	0.738	0.093	0.692	0.741	0.753
PGCL [1]	CVPR23	0.057	0.752	0.850	0.801	0.116	0.682	0.782	0.722	0.061	0.637	0.779	0.728	0.071	0.719	0.809	0.789
EPS [29]	TPAMI23	0.049	0.763	0.843	0.828	0.103	0.697	0.796	0.735	0.056	0.646	0.787	0.736	0.066	0.737	0.833	0.801
CoSOD [5]	WACV24	0.055	0.758	0.856	0.830	0.099	0.702	0.793	0.730	0.055	0.650	0.795	0.740	0.070	0.726	0.825	0.792
SEE [17]	TPAMI25	0.040	0.793	0.885	0.852	0.093	0.716	0.810	0.747	0.046	0.679	0.803	0.745	0.060	0.757	0.863	0.812
SCALER (Ours)	—	**0.039**	**0.802**	**0.886**	**0.857**	**0.090**	**0.727**	**0.832**	**0.755**	**0.044**	**0.688**	**0.822**	**0.760**	**0.057**	**0.771**	**0.866**	**0.822**
1/16 Labeled Training Data, PVT V2, SAM3																	
SAM3-S [6]	Arxiv 26	0.148	0.646	0.684	0.661	0.143	0.628	0.669	0.657	0.076	0.689	0.758	0.745	0.091	0.702	0.746	0.758
CoSOD [5]	WACV24	0.053	0.780	0.873	0.838	0.089	0.742	0.825	0.753	0.053	0.674	0.807	0.766	0.062	0.755	0.845	0.804
SEE [17]	TPAMI25	0.036	0.808	0.893	0.850	0.085	0.738	0.823	0.760	0.043	0.713	0.825	0.771	0.052	0.786	0.883	0.834
SCALER (Ours)	—	**0.031**	**0.828**	**0.905**	**0.874**	**0.075**	**0.757**	**0.854**	**0.779**	**0.034**	**0.733**	**0.849**	**0.790**	**0.045**	**0.812**	**0.906**	**0.847**

Table 3: Results on PIS and TOD of the WSCOS task with point supervision.

Methods	Polyp Image Segmentation (PIS)												Transparent Object Detection (TOD)							
	CVC-ColonDB				*ETIS*				*Kvasir*				*GDD*				*GSD*			
	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
Point Supervision, ResNet50, SAM																				
SAM [27]	0.479	0.343	0.419	0.427	0.429	0.439	0.512	0.503	0.320	0.545	0.564	0.582	0.245	0.512	0.530	0.551	0.266	0.473	0.501	0.514
SAM-Wp [7]	0.146	0.612	0.690	0.683	0.125	0.650	0.731	0.728	0.093	0.818	0.823	0.815	0.143	0.691	0.733	0.636	0.137	0.715	0.752	0.681
TEL [30]	0.089	0.669	0.743	0.761	0.083	0.639	0.776	0.726	0.091	0.810	0.826	0.804	0.230	0.640	0.586	0.536	0.275	0.571	0.501	0.495
SCOD [23]	0.077	0.691	0.795	0.802	0.071	0.664	0.802	0.766	0.071	0.853	0.877	0.836	0.146	0.801	0.778	0.723	0.154	0.743	0.751	0.710
WS-SAM [16]	0.043	0.721	0.839	0.816	0.037	0.694	0.849	0.797	0.046	0.878	0.917	0.877	0.078	0.858	0.863	0.775	0.089	0.839	0.841	0.764
SEE [17]	0.042	0.726	0.842	0.819	0.035	0.705	0.857	0.800	0.045	0.883	0.918	0.879	0.073	0.867	0.871	0.777	0.081	0.846	0.849	0.768
SCALER (Ours)	**0.041**	**0.735**	**0.849**	**0.824**	**0.034**	**0.715**	**0.868**	**0.806**	**0.043**	**0.904**	**0.930**	**0.887**	**0.071**	**0.877**	**0.880**	**0.782**	**0.078**	**0.857**	**0.859**	**0.770**

4 Experiments

Implementation details. We implement our method in PyTorch using two RTX 6000 Ada GPUs. Following [16], we adopt ResNet50 pre-trained on ImageNet [10] as the default backbone. All input images are resized to 352×352. The Adam optimizer is used with momentum parameters $(0.9, 0.999)$, batch size 12, and initial learning rate 0.0001 reduced by 0.1 every 30 epochs. The number of augmentation views K is set to 12. Our segmenter follows the same architecture as WS-SAM [16] for fair comparison. We fine-tune SAM using SAM-Adapter [7], but other adaptation strategies and foundation models (*e.g.*, SAM2 and SAM3) can be seamlessly integrated. Stronger foundation models and more effective fine-tuning strategies yield higher-quality pseudo-labels, providing further gains for both the segmenter and the foundation model via Phase II. Both models are first initialized on the available labeled data for robust optimization.

4.1 Comparative Evaluation

Following [16], we report results with point annotations under weak supervision. For COD, we also assess performance with scribble annotations provided by [13].
Camouflaged object detection. Tables 1 and 2 show that our SCALER framework outperforms competing methods in both weakly supervised and semi-

Table 4: Results on PIS and TOD of the SSCOS task.

Methods	Polyp Image Segmentation (PIS)												Transparent Object Detection (TOD)							
	CVC-ColonDB				*ETIS*				*Kvasir*				*GDD*				*GSD*			
	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
1/8 Labeled Training Data, ResNet50, SAM																				
SAM [27]	0.479	0.343	0.419	0.427	0.429	0.439	0.512	0.503	0.320	0.545	0.564	0.582	0.245	0.512	0.530	0.551	0.266	0.473	0.501	0.514
SAM-S [7]	0.185	0.517	0.627	0.661	0.172	0.558	0.706	0.682	0.131	0.738	0.772	0.783	0.185	0.674	0.703	0.618	0.177	0.688	0.724	0.652
PGCL [1]	0.068	0.629	0.743	0.749	0.057	0.623	0.771	0.764	0.070	0.838	0.869	0.855	0.097	0.819	0.823	0.747	0.095	0.817	0.804	0.741
EPS [29]	0.060	0.645	0.758	0.756	0.053	0.642	0.786	0.772	0.063	0.847	0.883	0.858	0.085	0.837	0.842	0.759	0.089	0.829	0.817	0.754
CoSOD [5]	0.057	0.657	0.776	0.763	0.047	0.639	0.802	0.780	0.056	0.859	0.886	0.863	0.088	0.831	0.835	0.756	0.091	0.826	0.813	0.749
SEE [17]	0.045	0.706	0.828	0.801	0.039	0.692	0.832	0.792	0.051	0.867	0.905	0.869	0.076	0.846	0.855	0.768	0.085	0.835	0.832	0.763
SCALER (Ours)	**0.043**	**0.726**	**0.841**	**0.823**	**0.038**	**0.701**	**0.851**	**0.798**	**0.050**	**0.873**	**0.914**	**0.879**	**0.073**	**0.854**	**0.864**	**0.781**	**0.083**	**0.846**	**0.845**	**0.777**
1/16 Labeled Training Data, ResNet50, SAM																				
SAM [27]	0.479	0.343	0.419	0.427	0.429	0.439	0.512	0.503	0.320	0.545	0.564	0.582	0.245	0.512	0.530	0.551	0.266	0.473	0.501	0.514
SAM-S [7]	0.201	0.471	0.593	0.643	0.183	0.533	0.677	0.660	0.145	0.712	0.753	0.762	0.203	0.657	0.672	0.604	0.193	0.654	0.684	0.627
PGCL [1]	0.075	0.608	0.719	0.745	0.063	0.614	0.753	0.754	0.078	0.836	0.845	0.835	0.108	0.794	0.803	0.735	0.108	0.800	0.788	0.718
EPS [29]	0.065	0.631	0.746	0.750	0.057	0.631	0.774	0.755	0.068	0.835	0.863	0.846	0.089	0.818	0.825	0.750	0.097	0.812	0.803	0.737
CoSOD [5]	0.063	0.638	0.754	0.755	0.050	0.632	0.781	0.767	0.066	0.840	0.871	0.852	0.095	0.813	0.816	0.746	0.105	0.793	0.792	0.732
SEE [17]	0.050	0.693	0.812	0.795	0.043	0.674	0.811	0.784	0.057	0.858	0.893	0.861	0.080	0.827	0.831	0.759	0.092	**0.824**	0.817	0.750
SCALER (Ours)	**0.048**	**0.698**	**0.821**	**0.806**	**0.041**	**0.686**	**0.822**	**0.799**	**0.056**	**0.878**	**0.899**	**0.885**	**0.078**	**0.851**	**0.848**	**0.767**	**0.082**	0.819	**0.831**	**0.768**

Table 5: Ablation study in COD on *COD10K* with the scribble supervision.

Metrics	Phase I					Phase II					SCALER
	w/o PL_G	w/o $E(\cdot)$	w/o $U(\cdot)$	$R_1(\cdot)\rightarrow R(\cdot)$	$R_2(\cdot)\rightarrow R(\cdot)$	w/o Phase II	w/o $\mathcal{L}_{ai}$	$\mathcal{L}_{ai}^1\rightarrow\mathcal{L}_{ai}$	w/o $\mathcal{L}_{nr}$	$\mathcal{L}_{nr}^1\rightarrow\mathcal{L}_{nr}$	(Ours)
$M\downarrow$	0.039	0.036	0.038	0.037	0.035	0.037	0.036	0.039	0.036	0.035	0.034
$F_\beta\uparrow$	0.732	0.724	0.721	0.728	0.722	0.723	0.733	0.730	0.729	0.734	0.737
$E_\phi\uparrow$	0.863	0.870	0.869	0.866	0.868	0.860	0.865	0.869	0.870	0.879	0.885
$S_\alpha\uparrow$	0.804	0.805	0.813	0.806	0.808	0.805	0.812	0.804	0.807	0.809	0.814

Table 6: Ablation on design choices. Columns: (3-6) threshold (τ_h, τ_l) variants, (8) Curriculum Learning, (9) SCALER default $(0.8, 0.2)$.

Metrics	Strong-weak Aug.	w/o R_b	(0.8,0.1)	(0.8,0.3)	(0.7,0.2)	(0.9,0.2)	Dynamic Top-k	Curriculum	SCALER
$M\downarrow$	0.037	0.036	0.035	0.036	0.036	0.035	0.034	0.035	0.034
$F_\beta\uparrow$	0.727	0.725	0.730	0.725	0.729	0.731	0.741	0.738	0.737
$E_\phi\uparrow$	0.870	0.868	0.876	0.871	0.874	0.877	0.885	0.881	0.885
$S_\alpha\uparrow$	0.806	0.804	0.810	0.805	0.808	0.811	0.818	0.815	0.814

supervised settings. Although fine-tuning SAM via SAMadapter [7] under point (SAM-$\mathrm{W_P}$), scribble (SAM-$\mathrm{W_S}$), and semi-supervised configurations yields performance improvements, these variants still underperform relative to SCALER. Furthermore, SCALER surpasses the previously proposed WS-SAM [16] and SEE [17] frameworks across all evaluation protocols. For example, SCALER generally outperforms the second-best method, SEE, in 4.27% and 5.72% in scribble and point supervision using PVT V2 and SAM3. This demonstrates our SCALER's effectiveness in producing higher-quality pseudo-labels. Fig. 5 provides qualitative comparisons under point supervision, further illustrating the advantages of our alternating optimization strategy.

Polyp image segmentation. Polyps often exhibit strong visual similarity to surrounding tissue, making PIS particularly challenging. As shown in Tables 3 and 4, our SCALER framework significantly outperforms the next best method, SEE, under point supervision. Both SAM and SAM-$\mathrm{W_P}$ perform poorly on this task, underscoring their limitations in handling concealed polyps. Fig. 5 illustrates that our approach more accurately localizes hidden polyps and produces segmentation boundaries that closely match the true lesion contours.

Transparent object detection. TOD is essential for enhancing vision systems in robotics, yet it remains challenging for the minimal visual contrast of objects. As shown in Tables 3 and 4, our SCALER surpasses current state-of-the-art methods under both weakly and semi-supervised settings. Fig. 5 illustrates that our method not only localizes transparent objects with high accuracy but also refines their segmentation boundaries and suppresses background clutter.

Table 7: Isolating Phase II's contribution. "+" denotes adding our Phase II to the base method's unchanged pipeline (retained as their own Phase I).

Metrics	Scribble, ResNet50, SAM				Scribble, PVT, SAM3				Point, ResNet50, SAM				Point, PVT, SAM3			
	WS-SAM	WS-SAM+	SEE	SEE+	WS-SAM	WS-SAM+	SEE	SEE+	WS-SAM	WS-SAM+	SEE	SEE+	WS-SAM	WS-SAM+	SEE	SEE+
$M \downarrow$	0.038	0.036	0.036	0.035	0.032	0.027	0.030	0.026	0.039	0.038	0.038	0.037	0.035	0.030	0.033	0.029
$F_\beta \uparrow$	0.719	0.728	0.729	0.738	0.751	0.775	0.760	0.773	0.698	0.710	0.706	0.718	0.724	0.750	0.735	0.758
$E_\phi \uparrow$	0.878	0.887	0.883	0.886	0.893	0.913	0.902	0.914	0.856	0.870	0.862	0.868	0.865	0.882	0.866	0.889
$S_\alpha \uparrow$	0.803	0.809	0.807	0.813	0.827	0.838	0.832	0.843	0.790	0.798	0.796	0.804	0.808	0.817	0.814	0.822

Table 8: Foundation model performance under different training strategies. "SAM+SAM" denotes homogeneous co-training with two identical SAM copies.

Metrics	Scribble, ResNet50, SAM					Scribble, PVT, SAM3				
	SAM	Adapter [7]	Conv-LoRA [51]	SAM+SAM	SCALER	SAM3	Adapter [6]	Conv-LoRA	SAM3+SAM3	SCALER
$M \downarrow$	0.093	0.049	0.046	0.038	**0.027**	0.088	0.047	0.044	0.030	**0.019**
$F_\beta \uparrow$	0.673	0.712	0.720	0.738	**0.765**	0.685	0.717	0.725	0.782	**0.853**
$E_\phi \uparrow$	0.737	0.833	0.842	0.870	**0.912**	0.748	0.836	0.845	0.903	**0.939**
$S_\alpha \uparrow$	0.730	0.770	0.778	0.800	**0.831**	0.738	0.773	0.781	0.826	**0.858**
Train. time (h)	—	12.4	18.2	28.6	22.2	—	14.8	21.5	38.5	31.8

4.2 Ablation Study

Experiments are conducted on the *COD10K* dataset with scribble supervision and ResNet50 backbone in Secs. 4.2 and 4.3.

Effectiveness in Phase I. As shown in Tab. 5, in Phase I, we observe performance degradation upon removing the fused pseudo-label PL_F, the entropy-based image-level weighting mechanism $E(\cdot)$, or the uncertainty-based pixel-level weighting mechanism $U(\cdot)$. Besides, omitting our specialized refinement procedures for difficult samples (denoted $R_1(\cdot)$) and simple samples (denoted $R_2(\cdot)$) also leads to a noticeable performance drop.

Effectiveness in Phase II. As shown in Tab. 5, removing Phase II entirely (*i.e.*, running only Stage 1 and 2) reverts SCALER to a one-way framework. This "w/o Phase II" ablation shows a significant performance drop, confirming that the mutual enhancement from Phase II is critical. We further evaluate the proposed augmentation invariance loss L_{ai} and noise resistance loss L_{nr}. Specifically, using an alternative L_{ai}^1, which imposes invariance between two weak augmentations only, or employing L_{nr}^1, which directly applies the refinement strategy $R(\cdot)$ from Phase I, results in decreased performance. This confirms that our proposed loss functions are better tailored for effectively fine-tuning SAM.

Design choices. Tab. 6 examines alternative design decisions in Phase II. (1) *Augmentation strategy*: unlike the strong-weak paradigm common in semi-supervised learning [1,37], weak-weak augmentation yields superior performance, as strong geometric or color distortions may disrupt the subtle foreground-background visual consistency, introducing noise that hinders representation learning. (2) *Basic loss*: removing $R_b(\cdot)$ causes a comparable drop, confirming that the basic supervised signal stabilizes training. (3) *Threshold sensitivity*: varying (τ_h, τ_l) across four combinations from (0.7, 0.1) to (0.9, 0.3) shows robust performance, justifying the default (0.8, 0.2). (4) *Adaptive alternatives*: "Dynamic Top-k"(selecting the top-k% most confident pixels with k linearly increasing) and "Curriculum Learning" (progressively raising τ_h) achieve competitive results on individual metrics but do not consistently outperform the fixed-threshold design across all four metrics, suggesting that the piecewise formulation offers a favorable simplicity-effectiveness trade-off.

Table 9: Comparison with co-training methods on *COD10K* (scribble). "Ho." uses two identical segmenters (Seg+Seg). "He." uses the same heterogeneous architecture as SCALER (Seg+SAM) but with the co-training method's symmetric loss. "+" employs SCALER's asymmetric optimization rules on the heterogeneous architecture.

Metrics	Co-teach (Ho.) [14]	Co-teach (He.)	Co-teach+	CPS (Ho.) [8]	CPS (He.)	CPS+	UCMT (Ho.) [36]	UCMT (He.)	UCMT+	SCALER
$M \downarrow$	0.051	0.043	0.038	0.045	0.040	0.036	0.043	0.038	0.035	**0.034**
$F_\beta \uparrow$	0.692	0.710	0.721	0.704	0.716	0.728	0.711	0.722	0.732	**0.737**
$E_\phi \uparrow$	0.842	0.858	0.871	0.853	0.864	0.876	0.857	0.869	0.879	**0.885**
$S_\alpha \uparrow$	0.778	0.792	0.802	0.786	0.796	0.807	0.790	0.800	0.811	**0.814**

Table 10: Cross-task generalization, with UniMatch [46] and S^{3}D [49] for semi-supervised semantic segmentation and infrared small target detection, and SAPNet [40] and EBSA [52] for weak-supervised instance segmentation and shadow detection.

Tasks (Datasets, Settings, Metrics)	Base	+ WS-SAM	+ SEE	+ SCALER
Sem. Seg. (VOC [11], 1/4, mIoU ↑)	76.8	76.2	76.4	77.9
Inst. Seg. (COCO [31], Point, mAP ↑)	34.6	33.0	33.5	35.8
IRSTD (SIRST [9], 1/16, IoU ↑)	84.6	80.5	84.3	88.2
Shadow (SBU [39], Scribble, BER ↓)	3.02	3.26	2.92	2.37

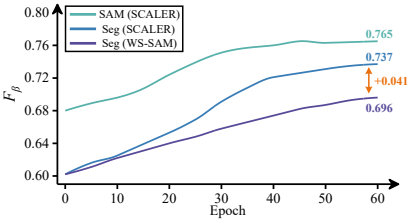

Fig. 6: Training dynamics and convergence analysis.

4.3 Further Analysis

Isolating Phase II's contribution. A key question is whether SCALER's gains stem primarily from Phase I or from the mutual learning loop. Tab. 7 addresses this by applying *only* Phase II on top of each base method's unchanged pipeline across all four settings. Phase II consistently improves every baseline and the improvements are notably amplified with stronger backbones, such as "Scribble, PVT, SAM3". Critically, all improvements are achieved without modifying any base method's Phase I, confirming the effect of Phase II.

Foundation model performance. Tab. 8 compares five training strategies for the foundation model across two settings. SCALER's heterogeneous pairing achieves substantially higher performance with shorter training time than the homogeneous training manner. This is because two identical SAMs share correlated error patterns on concealed objects, limiting mutual teaching; in contrast, the segmenter and SAM possess complementary failure modes, enabling each to correct the other's blind spots more effectively.

Comparison with co-training methods. We compare three co-training methods under three configurations (Tab. 9): (i) Ho. (homogeneous Seg+Seg), (ii) He. (heterogeneous Seg+SAM with a symmetric loss), and (iii) + (heterogeneous Seg+SAM with SCALER's asymmetric Phase I/Phase II rules). Three trends are observed. First, replacing a second segmenter with SAM (Ho.→He.) improves all methods, confirming the benefit of architectural complementarity. Second, applying SCALER's asymmetric optimization (He.→+) yields further gains, showing that heterogeneous architectures alone with a symmetric loss are insufficient and verifying the superiority of our SCALER.

Cross-task generalization. To evaluate whether SCALER's Phase II generalizes beyond COS, we apply it to four additional label-deficient segmentation tasks with representative specialist methods. As shown in Tab. 10, our SCALER can better enhance the performance of the segmenter in each case.

Training dynamics. Fig. 6 plots F_β on COD10K during 60-epoch training, where Phase I and Phase II alternate every epoch. Both models are initialized

Table 11: Computational cost comparison (RTX 6000 Ada, 352×352). SCALER trains both models but converges faster. At inference, only Seg_S is used.

Method	Train. Memory	Train. Time/Ep.	Epochs	Total train.	Params	FLOPs	Inf. Mem.	Inf. Time	$F_\beta \uparrow$
WS-SAM [16]	12.23G	13.67min	120	27.34h	31.12M	40.06G	2.31G	12.45ms	0.719
SEE [17]	15.78G	19.73min	100	33.07h	41.09M	49.64G	2.92G	15.33ms	0.729
SCALER (Seg)	16.37G	22.18min	60	22.18h	31.12M	40.06G	2.31G	12.45ms	0.737

via labeled-data pretraining. Two convergence patterns emerge: SAM converges earlier (∼epoch 45) as it only requires domain adaptation via Phase II; the segmenter, learning from a weaker initialization, converges later (∼epoch 55). In contrast, the WS-SAM baseline (the same segmenter) has *not* converged by epoch 60 (F_β: 0.696, still rising). The smooth, oscillation-free convergence empirically supports our theoretical guarantee (Sec. 3.4).

Computational cost. Tab. 11 compares computational costs. SCALER has higher per-epoch cost due to updating both models, but converges in only 60 epochs, yielding the lowest total training time. At inference, only Seg_S is deployed, so SCALER matches WS-SAM in inference cost while achieving the highest F_β. Note that we also bring a better foundation model (in Tab. 8). This further validates the effectiveness of our framework.

5 Discussions

Asymmetric collaboration as a general paradigm. SCALER reveals that a lightweight, task-specific segmenter can guide a large foundation model under label-deficient conditions, provided the two models receive asymmetric, bias-aware optimization. Our experiments (Tabs. 7 to 10) confirm this principle generalizes to various modules and tasks, suggesting a broader paradigm for other label-scarce domains (*e.g.*, medical imaging [25,26]) where knowledge should be exchanged reciprocally rather than through one-way distillation [16,17].

When does SCALER fail? We identify three scenarios. (1) *Cold-start failure*: when both models completely fail to localize a target, mutual learning cannot bootstrap; the labeled-data initialization mitigates this by establishing minimum competence. (2) *Systematic co-failure*: when both models share the same failure mode, mutual learning reinforces the error; SCALER's heterogeneous architecture reduces this risk vs. homogeneous co-training (Tab. 9), and entropy-based weighting further down-weights uncertain samples. (3) *Capacity mismatch*: when the domain gap exceeds SAM's adapter capacity (*e.g.*, electron microscopy), Phase II yields diminishing returns and SCALER gracefully degrades to a Phase I-only baseline. The training dynamics (Fig. 6) serve as a diagnostic: a flat SAM curve signals insufficient adapter capacity.

6 Conclusions

We proposed SCALER, a framework for LDCOS that enables asymmetric collaborative learning between a mean-teacher segmenter and a learnable SAM. Returning to our introductory questions: (1) consistency constraints and SAM-based supervision can indeed be jointly leveraged, as Phase I's dual-source pseudo-label refinement outperforms either alone; (2) the segmenter can guide

SAM through Phase II's augmentation invariance and noise resistance losses, yielding mutual improvement. The key insight is that this mutual enhancement requires *asymmetric* optimization tailored to each model's inductive bias, going beyond standard co-training or one-way distillation.

Acknowledgements

This work was supported in part by the Foundation Fighting Blindness (BR-CL-0621-0812-DUKE and PPA-1224-0890-DUKE) and a Research to Prevent Blindness Unrestricted Grant to Duke University.

References

1. Basak, H., Yin, Z.: Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In: CVPR. pp. 19786–19797 (2023) 4, 10, 11, 12
2. Bertsekas, D.P., et al.: Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. Optimization for Machine Learning **2010**(1-38), 3 (2011) 9
3. Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.: Clarke subgradients of stratifiable functions. SIAM Journal on Optimization **18**(2), 556–572 (2007) 9
4. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Mathematical Programming **146**(1), 459–494 (2014) 9
5. Chakraborty, S., Naha, S., Bastan, M., Samaras, D., et al.: Unsupervised and semi-supervised co-salient object detection via segmentation frequency statistics. In: WACV. pp. 332–342 (2024) 4, 10, 11
6. Chen, T., Cao, R., Yu, X., Zhu, L., Ding, C., Ji, D., Chen, C., Zhu, Q., Xu, C., Mao, P., et al.: Sam3-adapter: Efficient adaptation of segment anything 3 for camouflage object segmentation, shadow detection, and medical image segmentation. arXiv preprint arXiv:2511.19425 (2025) 9, 10, 12
7. Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S.: Sam-adapter: Adapting segment anything in underperformed scenes. In: ICCV. pp. 3367–3375 (2023) 3, 7, 9, 10, 11, 12
8. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2613–2622 (2021) 2, 3, 4, 13
9. Dai, Y., Wu, Y., Zhou, F., Barnard, K.: Asymmetric contextual modulation for infrared small target detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 950–959 (2021) 13
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009) 10
11. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010) 13
12. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J.: Camouflaged object detection. In: CVPR. pp. 2777–2787 (2020) 2, 4

13. Gao, S., Zhang, W., Wang, Y., Guo, Q., Zhang, C., He, Y., Zhang, W.: Weakly-supervised salient object detection using point supervison. In: AAAI. vol. 36, pp. 670–678 (2022) 10
14. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in neural information processing systems **31** (2018) 2, 3, 4, 13
15. He, C., Li, K., Zhang, Y., Tang, L., Zhang, Y.: Camouflaged object detection with feature decomposition and edge reconstruction. In: CVPR. pp. 22046–22055 (2023) 4
16. He, C., Li, K., Zhang, Y., Xu, G., Tang, L.: Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. NeurIPS (2023) 2, 3, 4, 6, 9, 10, 11, 14
17. He, C., Li, K., Zhang, Y., Yang, Z., Pang, Y., Tang, L., Fang, C., Zhang, Y., Kong, L., Li, X., Farsiu, S.: Segment concealed objects with incomplete supervision. IEEE Trans. Pattern Anal. Mach. Intell. (2025) 2, 3, 4, 6, 9, 10, 11, 14
18. He, C., Li, K., Zhang, Y., Zhang, Y., Guo, Z., Li, X.: Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. ICLR (2024) 4
19. He, C., Zhang, R., Xiao, F., Fang, C., Tang, L., Zhang, Y., Farsiu, S.: Unfoldir: Rethinking deep unfolding network in illumination degradation image restoration. arXiv preprint arXiv:2505.06683 (2025) 2
20. He, C., Zhang, R., Xiao, F., Fang, C., Tang, L., Zhang, Y.: Run: Reversible unfolding network for concealed object segmentation. ICML (2025) 4
21. He, C., Zhang, R., Xiao, F., Zhang, D., Cao, Z., Farsiu, S.: Refining context-entangled content segmentation via curriculum selection and anti-curriculum promotion. In: ICML (2026) 2
22. He, C., Zhang, R., Zhang, D., Fang, C., Tang, L., Feng, J., Xiao, F., Farsiu, S.: Ride: Retinex-informed decoupling for exposing concealed objects. arXiv preprint arXiv:2605.15450 (2026) 2
23. He, R., Dong, Q., Lin, J.: Weakly-supervised camouflaged object detection with scribble. AAAI (2023) 2, 4, 9, 10
24. Hu, J., Lin, J., Gong, S.: Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In: AAAI. vol. 38, pp. 12511–12518 (2024) 4
25. Ji, G.P., Fan, D.P., Xu, P., Cheng, M.M., Zhou, B.: Sam struggles in concealed scenes". arXiv preprint arXiv:2304.06022 (2023) 14
26. Ji, W., Li, J., Bi, Q., Li, W., Cheng, L.: Segment anything is not always perfect: An investigation of sam on different real-world applications. arXiv preprint arXiv:2304.05750 (2023) 14
27. Kirillov, A., Mintun, E., Ravi, N., Mao, H.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 2, 4, 9, 10, 11
28. Lai, X., Yang, Z., Hu, J., Zhang, S., Cao, L.: Camoteacher: Dual-rotation consistency learning for semi-supervised camouflaged object detection. arXiv preprint arXiv:2408.08050 (2024) 2, 4
29. Lee, M., Lee, S., Lee, J., Shim, H.: Saliency as pseudo-pixel supervision for weakly and semi-supervised semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **45**(10), 12341–12357 (2023) 4, 10, 11
30. Liang, Z., Wang, T., Zhang, X., Sun, J., Shen, J.: Tree energy loss: Towards sparsely annotated semantic segmentation. In: CVPR. pp. 16907–16916 (2022) 9, 10

31. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 13
32. Nicolas, C., Laura, G., Yuan-Ting, H., Shoubhik, D., Ronghang, H., , et al.: Sam 3: Segment anything with concepts (2025), https://arxiv.org/abs/2511.16719 2
33. Niu, Y., Yang, L., Xu, R., Li, Y., Chen, Y.: Minet: Weakly-supervised camouflaged object detection through mutual interaction between region and edge cues. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 6316–6325 (2024) 4
34. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024) 2
35. Ren, G., Lazarou, M., Yuan, J., Stathaki, T.: Towards automated polyp segmentation using weakly-and semi-supervised learning and deformable transformers. In: CVPR. pp. 4355–4364 (2023) 4
36. Shen, Z., Cao, P., Yang, H., Liu, X., Yang, J., Zaiane, O.R.: Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. arXiv preprint arXiv:2301.04465 (2023) 2, 4, 13
37. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. NeurIPS **30** (2017) 4, 12
38. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of optimization theory and applications **109**(3), 475–494 (2001) 8
39. Vicente, T.F.Y., Hou, L., Yu, C.P., Hoai, M., Samaras, D.: Large-scale training of shadow detectors with noisily-annotated shadow examples. In: European conference on computer vision. pp. 816–832. Springer (2016) 13
40. Wei, Z., Chen, P., Yu, X., Li, G., Jiao, J., Han, Z.: Semantic-aware sam for point-prompted instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3585–3594 (June 2024) 13
41. Xiao, F., Feng, J., Hu, P., Zhang, D., Xu, L., Qin, G., Li, L., He, C., Farsiu, S.: Qualiteacher: Quality-conditioned pseudo-labeling for real-world image restoration. ECCV (2026) 2
42. Xiao, F., Hu, P., Xu, L., Guo, X., Qin, G., Shen, Y., Fang, C., Zhang, R., He, C., Farsiu, S.: Beyond ground-truth: Leveraging image quality priors for real-world image restoration. CVPR (2026) 2
43. Xiao, F., Hu, S., Shen, Y., He, C.: A survey of camouflaged object detection and beyond. CAAI AIR (2024) 2
44. Xiao, F., Zhang, P., He, C.: Concealed object segmentation with hierarchical coherence modeling. In: CAAI ICAI. pp. 16–27 (2023) 2
45. Xu, Y., Yin, W.: A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. SIAM Journal on imaging sciences **6**(3), 1758–1789 (2013) 8
46. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7236–7246 (June 2023) 13
47. Yang, Z., Choy, K., Farsiu, S.: Spatial coherence loss: All objects matter in salient and camouflaged object detection. Pattern Recognition p. 112798 (2025) 4
48. Yin, C., Yang, K., Li, J., Li, X.: Mutual iterative refinement network for scribble-supervised camouflaged object detection. IEEE Transactions on Image Processing **34**, 7347–7361 (2025) 4

49. Zhang, M., Shang, W., Gao, F., Zhang, Q., Lu, F., Zhang, J.: Semi-supervised infrared small target detection with thermodynamic-inspired uneven perturbation and confidence adaptation. Proceedings of the AAAI Conference on Artificial Intelligence **39**(10), 10013–10021 (Apr 2025). https://doi.org/10.1609/aaai.v39i10.33086, https://ojs.aaai.org/index.php/AAAI/article/view/33086 13
50. Zheng, Y., Zhong, B., Liang, Q., Zhang, S., Li, G., Li, X., Ji, R.: Towards universal modal tracking with online dense temporal token learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 2
51. Zhong, Z., Tang, Z., He, T., Fang, H., Yuan, C.: Convolution meets lora: Parameter efficient finetuning for segment anything model. arXiv preprint arXiv:2401.17868 (2024) 12
52. Zhou, K., Fang, J., Wei, D., Wu, W., Hu, R.: Exploring better sparsely annotated shadow detection. Neural Networks **181**, 106827 (2025). https://doi.org/https://doi.org/10.1016/j.neunet.2024.106827, https://www.sciencedirect.com/science/article/pii/S0893608024007512 13