# SEED: SUBJECT-TO-EDGE EXPLORATION AND DISTILLATION FOR UNSUPERVISED CAMOUFLAGED OBJECT DETECTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Unsupervised Camouflaged Object Detection (UCOD) aims to identify concealed objects without relying on pixel-level annotations. Mainstream methods rely on prior-based pseudo-labels that often misclassify background noise as foreground under high object–background similarity. Recent approaches suppress noise via iterative pseudo-label optimization or uncertainty-aware strategies. However, they face a trade-off between noise filtering and detail preservation: conservative filtering fails to remove noise that disrupts subject localization, whereas aggressive filtering sacrifices edge details for subject purification. To address these issues, we decompose UCOD into a subject localization with distillation phase and a detail exploration phase. Accordingly, we propose a **S**ubject-to-**E**dge **E**xploration and **D**istillation (SEED) method based on the Teacher-Student framework, which comprises three modules: the Subject Teacher Initialization (STI) module, the Subject Consistency Distillation (SCD) module, and the Active Detail Exploration (ADE) module. For subject localization with distillation, STI establishes a purified camouflaged subject via dynamic pseudo-label, then SCD employs confidence-partitioned distillation to maintain subject consistency and mitigate confusion. For detail exploration, ADE leverages temperature rescaling to actively refine details for fine-grained prediction. Extensive experiments demonstrate that our method achieves state-of-the-art performance.

## 1 INTRODUCTION

Camouflaged Object Detection (COD) aims to identify objects visually concealed in their surroundings Fan et al. (2020a), which has significant applications in several important fields Fan et al. (2020b); Cannaday et al. (2023); Li et al. (2024). However, it is difficult to obtain large-scale human labels at the pixel-level for camouflaged objects due to their complex visual attributes (e.g., texture, color, and structural patterns Lai et al. (2024); Yan et al. (2025a)). Unlike fully-supervised methods, Unsupervised Camouflaged Object Detection (UCOD) aims to identify camouflaged objects without any pixel-level human annotation. As a result, UCOD has attracted increasing attention from researchers.

Mainstream UCOD methods typically rely on pseudo-labels generated from prior knowledge to provide supervision Siméoni et al. (2023); Zhang & Wu (2023); Shou et al. (2025); Du et al. (2025b); Yan et al. (2025a). Since camouflaged objects exhibit extremely high appearance similarity to the background, such pseudo-labels often contain substantial foreground-like background noise. To address this, recent UCOD methods focus mainly on noise filtering of pseudo-labels, e.g., by iterative optimization to update pseudo-labels Yan et al. (2025a) or by using uncertainty-aware strategies to localize high-confidence foreground regions Shou et al. (2025); Du et al. (2025b). Despite some success in noise suppression, the predictions of existing methods are often degraded and lack fine-grained details. As illustrated in fig. 1(a), existing methods tend to oscillate between two extremes: retaining background distractors due to under-filtering, or eroding object details due to over-suppression, which are shown in fig. 2. Consequently, they face a trade-off between noise filtering and detail preservation: conservative noise



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.

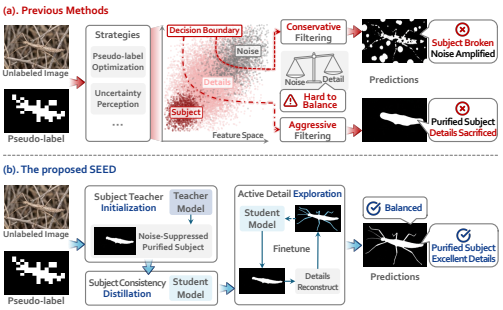



Figure 1: **Comparison between previous UCOD methods and SEED.** Existing methods suffer from the noise-detail trade-off, while SEED resolves it through subject distillation and detail exploration.

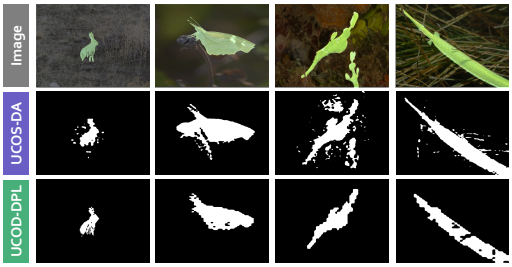



Figure 2: **Illustration of the limitations of existing UCOD methods.** UCOS-DA Zhang & Wu (2023) tends to introduce background noise due to conservative filtering, while UCOD-DPL Yan et al. (2025a) erodes object boundaries due to aggressive filtering.

filtering fails to remove noise that disrupts subject localization, whereas aggressive noise filtering achieves subject purification but sacrifices edge details and subtle structural details.

Instead of being trapped in the trade-off between noise filtering and detail preservation, we propose to decouple the UCOD task into subject localization with distillation and detail exploration, as illustrated in fig. 1(b). We adopt a coarse-to-fine strategy that localizes a purified camouflaged subject first and then explores fine boundary details under subject constraints. Consequently, we propose a Teacher–Student framework for **S**ubject-to-**E**dge **E**xploration and **D**istillation (SEED), where the teacher model localizes a purified camouflaged subject, and the student model progressively explores and refines fine details constrained by the camouflaged subject of the teacher model. The framework consists of three progressive modules: Subject Teacher Initialization (STI), Subject Consistency Distillation (SCD), and Active Detail Exploration (ADE). In the subject localization with distillation phase, the STI trains the teacher model with a dynamic pseudo-label to obtain a purified camouflaged subject. Next, the SCD partitions student predictions into high-confidence and confusion regions to apply a confidence-based region partition constraint. It enforces strict teacher-student consistency within high-confidence regions to stabilize the camouflaged subject, while employing our proposed confusion suppression strategy to mitigate background noise in uncertain regions. In the subsequent detail exploration phase, ADE extends confident subject predictions by mining object skeletons for topological cues. It then applies adaptive local temperature rescaling to sharpen predictive distributions, compelling the student model to recover subtle structural nuances and ultimately achieving both precise subject localization and fine-grained boundary preservation.

To validate the effectiveness of our proposed method, we carried out extensive experiments on four main COD benchmark datasets (CHAMELEON, CAMO, COD10K, and NC4K). In particular, our method improves the Mean Absolute Error (MAE) by 23.02% and the S-Measure by 2.47% compared to previous state-of-the-art UCOD methods.

Our main contributions are summarized as follows:

1. We propose a novel **S**ubject-to-**E**dge **E**xploration and **D**istillation (SEED) framework for the UCOD, which is decoupled into subject localization with distillation and detail exploration to resolve the trade-off between noise suppression and detail preservation. Extensive experiments demonstrate its outstanding performance.
2. In the subject localization with distillation phase, we introduce a STI module to train a teacher model with dynamic pseudo-label for purified camouflaged subject localization, and a SCD module to distill confidence-aware subject consistency into the student while suppressing background noise in uncertain regions.
3. In the detail exploration phase, we propose an ADE module that employs an adaptive local temperature rescaling mechanism to refine predictions and actively recover subtle boundary and structural details under purified subject constraints.

# 2 Related Works

## 2.1 Camouflaged Object Detection

Recent work on camouflaged object detection (COD) Zhang et al. (2025); Liu et al. (2024); Zheng et al. (2024) has explored unsupervised approaches (UCOD) that learn without pixel-level annotations. Existing UCOD methods seek to suppress background noise mistaken for foreground due to high object–background similarity. Existing UCOD methods can be viewed as different attempts to suppress noisy foreground cues caused by the high similarity between camouflaged objects and their surroundings. For example, UCOS-DA Zhang & Wu (2023) formulates UCOD as a source-free domain adaptation problem, where domain alignment helps reduce unreliable responses transferred from non-camouflaged source data. EASE Du et al. (2025a) introduces environmental region modeling to distinguish camouflaged targets from the background context, thus filtering foreground-like environmental distractors. More recent pseudo-label learning methods address the noise issue more directly. UCOD-DPL Yan et al. (2025a) suppresses noisy pseudo-labels through dynamic pseudo-label mixing, while EReCu Jiang et al. (2026) and Self-Anchored Liu et al. (2026) progressively evolve or refine pseudo-labels to reduce accumulated errors during training. However, existing methods typically optimize pseudo-labels globally within a coupled iterative framework. As a result, they struggle to balance noise filtering and detail preservation to camouflaged object boundaries, which are visually similar to the background. In contrast, our SEED adopts a decoupled training paradigm that separates subject localization from detail exploration. It first localizes a purified camouflaged subject and applies confidence-partitioned distillation to maintain subject consistency and mitigate confusion. Subsequent detail exploration uses temperature rescaling to recover edge details, addressing the trade-off between noise filtering and detail preservation.

## 2.2 Self-Supervised Learning and Self-Training

Self-supervised learning (SSL) learns visual representations using supervision derived from unlabeled images. Representative approaches include contrastive learning He et al. (2020); Chen et al. (2020), masked image modeling Bao et al. (2021); He et al. (2022); Xie et al. (2022), and teacher–student learning Grill et al. (2020). DINO and DINOv2 Caron et al. (2021); Oquab et al. (2023) learn through self-distillation across image views and capture object-level structure in their features. Beyond representation learning, self-supervised features also support unsupervised segmentation. TokenCut Wang et al. (2022) discovers foreground regions through normalized cuts on self-supervised feature graphs. FOUND Siméoni et al. (2023) uses background-derived pseudo-labels for self-training. In UCOD, SdalsNet Shou et al. (2025) employs self-distillation for attention localization and shifting. UCOD-DPL Yan et al. (2025a) dynamically mixes fixed-strategy pseudo-labels with teacher predictions for self-training.

Despite these advances, pseudo-label-based self-training can reinforce residual errors when noisy predictions are repeatedly reused as supervision, a risk associated with confirmation bias Arazo et al. (2020). This concern is relevant to UCOD, where high object–background similarity can cause background noise to be mislabeled as foreground Yan et al. (2025a). Reusing such mislabeled regions as training targets may preserve or amplify false foreground responses rather than correct them. SEED uses subject purification and consistency to improve the reliability of self-generated supervision, reducing the risk of noise reinforcement during detail exploration.

# 3 Methods

## 3.1 Overall Framework

The overall framework of our proposed SEED is shown in fig. 3, which comprises a pre-trained feature extractor and a Teacher-Student framework with a Subject Teacher Initialization (STI) module, a Subject Consistency Distillation (SCD) module, and an Active Detail Exploration (ADE) module. We decouple the UCOD into two phases: the subject localization with distillation phase and the detail exploration phase.

Given an unlabeled image $I_i^{in}$, we first utilize a pre-trained feature extractor to generate hierarchical semantic features $\{F_i^l\}_{l=1}^{L}$. In the subject localization with distillation phase, we use STI to initialize

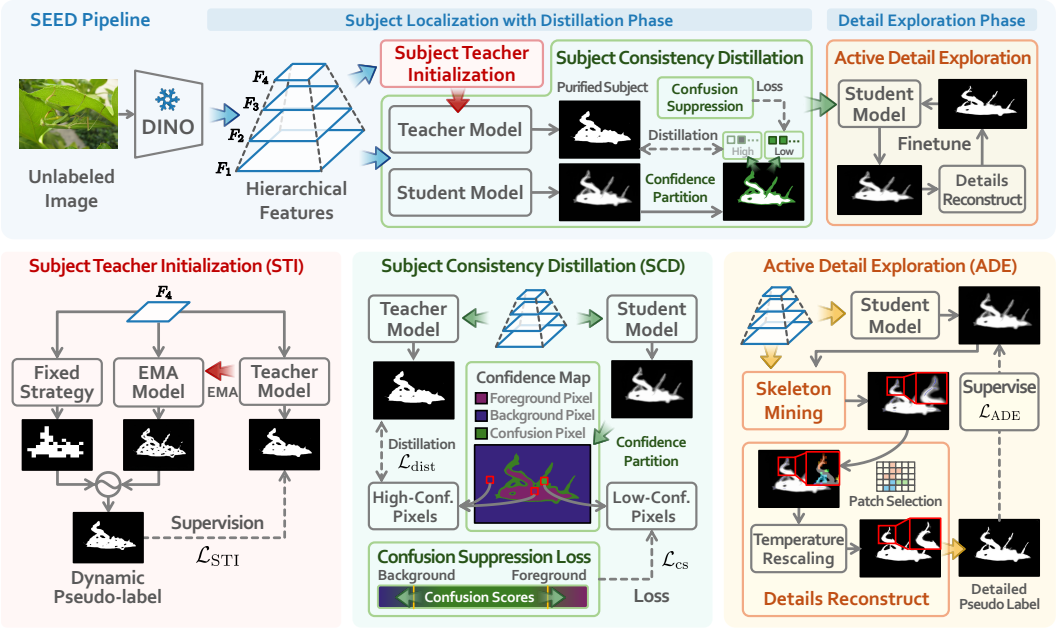



Figure 3: **The main framework of our proposed method.** The SEED contains a teacher-student framework, a Subject Teacher Initialization (STI) module, a Subject Consistency Distillation (SCD) module, and an Active Detail Exploration (ADE) module.

the teacher model $\mathcal{D}^T$, which can produce predictions with a purified camouflaged subject. Then, the SCD module performs confidence-partitioned distillation between the teacher and student models, transferring the purified camouflaged subject generated by the teacher model $\mathcal{D}^T$ to the student model $\mathcal{D}^S$. Specifically, SCD partitions the student prediction $\hat{P}_i^s$ into the high-confidence region and confusion region according to confidence scores, enforces subject consistency distillation on the high-confidence region to align with the teacher prediction $\hat{P}_i^t$, and applies a confusion suppression loss on the confusion region to mitigate residual noise.

In the detail exploration phase, ADE reconstructs details from the student model's predictions $\hat{P}_i^s$ and finetunes the student model using the reconstructed fine-grained pseudo-label $\hat{P}_i^{TR}$, enhancing its ability to recover fine-grained details. Specifically, the ADE module applies a Soft-Skeletonization operation on the student prediction $\hat{P}_i^s$ to mine the structural skeleton map, which serves as a topology-aware cue for exploring complex edges and subtle structure details. Based on the skeleton map, an adaptive local temperature rescaling mechanism is further introduced to refine the prediction and generate a fine-grained pseudo-label $\hat{P}_i^{TR}$ to fine-tune the student model.

During inference, only the student model is employed to generate predictions for input images.

### 3.2 Subject Teacher Initialization

Mainstream approaches typically rely on heuristic priors to produce initial pseudo-labels. However, as noted in previous work Yan et al. (2025a); Du et al. (2025b), such pseudo-labels often contain substantial noise due to the intrinsic similarity between camouflaged objects and the background. Training directly on such pseudo-labels will lead to the accumulation of incorrect knowledge. To provide a purified camouflaged subject for model training, we propose the Subject Teacher Initialization (STI) module.

First, given an unlabeled input image $I_i^{in}$, we first generate coarse pseudo-labels $\hat{P}_i^{fs}$ using a fixed strategy proposed by Siméoni et al. (2023). In parallel, we maintain an EMA model of the teacher model, whose parameters are kept frozen during training and are updated using the Exponential Moving Average (EMA) Tarvainen & Valpola (2017) of the teacher model's parameters:

$$\theta_t^{ema} = \alpha\theta_{t-1}^{ema} + (1-\alpha)\theta_t^T, \tag{1}$$

where $\alpha$ is the momentum coefficient and set to 0.99 following Yan et al. (2025a) and $t$ denotes the training step.

Then, we feed the last-layer features $F_i^L$ extracted from the pre-trained feature extractor into the teacher model and the EMA model to obtain teacher model prediction $\hat{P}_i^t$ and EMA model prediction $\hat{P}_i^{ema}$:

$$\hat{P}_i^{ema} = \sigma(\mathcal{D}^T(F_i^L; \theta_t^{ema})),\ \hat{P}_i^t = \sigma(\mathcal{D}^T(F_i^L; \theta_t^T)), \tag{2}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function and $\mathcal{D}^T(\cdot; \theta_t^{ema})$ represents the EMA model and $\mathcal{D}^T(\cdot; \theta_t^T)$ denotes the teacher model. Since the EMA model is uninformative in early training but becomes increasingly reliable as training proceeds, we follow the dynamic learning paradigm of UCOD-DPL Yan et al. (2025a) and construct a dynamic pseudo-label using a temporally scheduled linear-mixing strategy, where the weight grows linearly with the training step:

$$\begin{aligned} P_{dy}^{(t)} &= \mathrm{Linear}(\hat{P}_i^{ema}, \hat{P}_i^{fs}),\ \lambda_t = \frac{t}{T_{total}}, \\ \mathrm{Linear}(\hat{P}_i^{ema}, \hat{P}_i^{fs}) &= \lambda_t \cdot \hat{P}_i^{ema} + (1 - \lambda_t) \cdot \hat{P}_i^{fs}, \end{aligned} \tag{3}$$

where $\mathrm{Linear}(\cdot, \cdot)$ represents the time-dependent linearly fusion function, $t$ denotes the current training step and $T_{total}$ represents the total number of training steps. Then, we employ a binary cross-entropy loss $\mathcal{L}_{BCE}$ to supervise the teacher model:

$$\mathcal{L}_{\mathrm{STI}} = \mathcal{L}_{\mathrm{BCE}}(\hat{P}_i^t, P_{dy}^{(t)}). \tag{4}$$

### 3.3 Subject Consistency Distillation

Although the teacher model initialized by STI can produce purified camouflaged subjects, it often neglects complex edges and subtle structure details. A direct global distillation would propagate these neglects to the student model, leading to blurred edges and increased confusion with background noise. Therefore, we introduce the Subject Consistency Distillation (SCD) module to resolve this using a confidence-based partitioning strategy. Specifically, SCD imposes distinct loss constraints on high-confidence and confusion regions, effectively suppressing background noise confusion while ensuring subject consistency.

Specifically, based on the prediction certainty threshold $\tau_{\mathrm{ce}}$, we partition the student model's prediction $\hat{P}_i^s$ into a high-confidence region $M_i^{\mathrm{hc}}$ and a confusion region $M_i^{\mathrm{unc}}$:

$$\begin{aligned} M_i^{\mathrm{hc}} &= \mathbb{1}(\hat{P}_i^s > \tau_{\mathrm{ce}}) \cup \mathbb{1}(\hat{P}_i^s < (1 - \tau_{\mathrm{ce}})), \\ M_i^{\mathrm{unc}} &= 1 - M_i^{\mathrm{hc}}, \end{aligned} \tag{5}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function; we set $\tau_{\mathrm{ce}} = 0.7$ and a sensitivity analysis of $\tau_{ce}$ is provided in the supplementary material. The mask $M_i^{\mathrm{hc}}$ covers pixels where the student model exhibits high certainty for foreground or background. Conversely, $M_i^{\mathrm{unc}}$ denotes the pixel set in the confusion region, where pixel confidences fall between the two thresholds. This region typically corresponds to fine edges or camouflaged textures, often harboring ambiguous information that diverges from the teacher's prediction $\hat{P}_i^t$.

For pixels in the high-confidence region $M_i^{\mathrm{hc}}$, we impose a subject consistency constraint to distill the teacher's capability for producing purified subjects into the student model. We first compute the bounding box of the foreground from the student model prediction $\hat{P}_i^s$. The input image is then cropped using the expanded bounding box and fed back into the student model to generate a localized prediction $\hat{P}_i^{\mathrm{ROI}}$:

$$\hat{P}_i^{\mathrm{ROI}} = \mathrm{Paste}(\mathcal{D}^S(\mathrm{Crop}(I_i^{in}, \hat{P}_i^s))), \tag{6}$$

where the $\mathcal{D}^S$ denotes the student model. $\mathrm{Crop}(I, P)$ represents cropping the image $I$ using the bounding box predicted by prediction $P$, $\mathrm{Paste}(P)$ represent paste prediction $P$ to the original area.

The final supervision target $T_i$ is generated by linearly mixing the localized prediction $\hat{P}_i^{\mathrm{ROI}}$ and the teacher's prediction $\hat{P}_i^t$:

$$T_i = \mathrm{Linear}(\hat{P}_i^{\mathrm{ROI}}, \hat{P}_i^t)$$
$$\mathcal{L}_{dist} = \frac{\sum \left( M_i^{\mathrm{hc}} \odot \mathcal{L}_{Seg}(\hat{P}_i^s, T_i) \right)}{\sum M_i^{\mathrm{hc}}} \tag{7}$$

where $\mathcal{L}_{Seg}$ denotes the segmentation loss between the student prediction and the final supervision target, and $\mathrm{Linear}(\cdot,\cdot)$ follows the same definition as eq. (3).This strategy gradually fades out the teacher's supervision, allowing the student to surpass the teacher's coarse predictions and learn finer details later in training.

In contrast, for the confusion region $M_i^{\mathrm{unc}}$, we disregard the teacher's guidance to mitigate noise confusion. Instead, we aim to stimulate the student's intrinsic discriminative capability. To prevent the student model from being trapped in ambiguity, we introduce a warm-up mechanism to encourage decisive predictions. Specifically, we gather and sort the student's pixel-wise scores within the confusion region $M_i^{\mathrm{unc}}$. We then select the Top-$K\%$ pixels $\mathcal{S}_{top}$ and Bottom-$K\%$ pixels $\mathcal{S}_{bot}$, pushing these relatively certain pixels towards the high-confidence thresholds. This confusion suppression loss is formulated as:

$$\mathcal{L}_{cs} = \frac{1}{2K} \left( \sum_{p \in \mathcal{S}_{top}} \|p - \tau_{\mathrm{ce}}\|^2 + \sum_{p \in \mathcal{S}_{bot}} \|p - (1 - \tau_{\mathrm{ce}})\|^2 \right). \tag{8}$$

Finally, the total loss for the SCD module is the weighted sum of these components:

$$\mathcal{L}_{\mathrm{SCD}} = \mathcal{L}_{dist} + \beta \mathcal{L}_{cs}. \tag{9}$$

In our experiments, we empirically set $\beta = 1.0$, since the two loss terms are within the same numerical range.

### 3.4 Active Detail Exploration

To further recover fine-grained camouflaged object details within confusion regions following Subject Consistency Distillation, we propose the Active Detail Exploration (ADE) module. Given the extreme similarity between subtle object boundaries and background noise, direct differentiation is highly challenging; thus, we model potential target details using the topological structure of the prediction map. Specifically, ADE identifies potential object extension regions using skeleton structures and sharpens their detail responses via an adaptive local temperature rescaling mechanism. Unlike traditional skeletonization methods Zhang & Suen (1984) restricted to binarized masks, we adopt the Soft-Skeletonization operation Shit et al. (2021) to directly extract the structural skeleton map $M_i^{\mathrm{skel}}$ from the student model's prediction $\hat{P}_i^s$:

$$M_i^{\mathrm{skel}} = \mathrm{Skeletonize}(\hat{P}_i^s), \tag{10}$$

where the $\mathrm{Skeletonize}(\cdot)$ denoted Soft-Skeletonization operation. We view the structural skeleton as a reliable cue for topology and then perform morphological connectivity analysis to partition the image into a set of disjoint detail exploration regions $\Omega = \bigcup_{k=1}^{N_i} C_k$. Specifically, these regions are defined as the connected components of the extracted skeleton structures:

$$\{C_k\}_{k=1}^{N_i} = \mathrm{PatchSelect}(M_i^{\mathrm{skel}}), \tag{11}$$

where $\mathrm{PatchSelect}(\cdot)$ groups adjacent patches into disjoint regions following the connectivity of the skeleton, and $C_k$ denotes the $k$-th connected component.

Table 1: Comparison of our methods with recent COD methods.

| Methods | CHAMELEON (76) | | | | | CAMO-Test (250) | | | | | COD10K-Test (2,026) | | | | | NC4K (4,121) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{S}_m\uparrow$ | $\mathcal{F}_\beta^w\uparrow$ | $\mathcal{F}_\beta^m\uparrow$ | $\mathcal{E}_\phi^m\uparrow$ | $\mathcal{M}\downarrow$ | $\mathcal{S}_m\uparrow$ | $\mathcal{F}_\beta^w\uparrow$ | $\mathcal{F}_\beta^m\uparrow$ | $\mathcal{E}_\phi^m\uparrow$ | $\mathcal{M}\downarrow$ | $\mathcal{S}_m\uparrow$ | $\mathcal{F}_\beta^w\uparrow$ | $\mathcal{F}_\beta^m\uparrow$ | $\mathcal{E}_\phi^m\uparrow$ | $\mathcal{M}\downarrow$ | $\mathcal{S}_m\uparrow$ | $\mathcal{F}_\beta^w\uparrow$ | $\mathcal{F}_\beta^m\uparrow$ | $\mathcal{E}_\phi^m\uparrow$ | $\mathcal{M}\downarrow$ |
| ***Fully-Supervised Methods*** | | | | | | | | | | | | | | | | | | | | |
| SINetFan et al. (2020a) | .872 | .806 | .827 | .946 | .034 | .751 | .606 | .675 | .771 | .100 | .771 | .551 | .634 | .806 | .051 | .808 | .723 | .769 | .871 | .058 |
| BGNetSun et al. (2022) | .901 | .851 | .860 | .954 | .027 | .812 | .749 | .789 | .870 | .073 | .831 | .722 | .753 | .901 | .033 | .851 | .788 | .820 | .907 | .044 |
| ZoomNetPang et al. (2022) | .902 | .845 | .864 | .958 | .023 | .820 | .752 | .794 | .878 | .066 | .838 | .729 | .766 | .888 | .029 | .853 | .784 | .818 | .896 | .043 |
| SINetv2Fan et al. (2022) | .888 | .816 | .835 | .961 | .030 | .820 | .743 | .782 | .882 | .070 | .815 | .680 | .718 | .887 | .037 | .847 | .770 | .805 | .903 | .048 |
| HitNetHu et al. (2023) | **.921** | .897 | **.900** | .972 | .019 | .849 | .809 | **.831** | .906 | .055 | .871 | .806 | **.823** | .935 | .023 | .875 | .834 | **.853** | .926 | .037 |
| FSPNetHuang et al. (2023) | .908 | .851 | .867 | .965 | .023 | .856 | .799 | .830 | .899 | .050 | .851 | .735 | .769 | .895 | .026 | .879 | .816 | .843 | .915 | .035 |
| CDPLiu et al. (2024) | .865 | .- | .819 | .923 | .033 | .798 | - | .767 | .844 | .072 | .791 | - | .709 | .863 | .036 | .880 | - | .799 | .902 | .051 |
| FGSA-NetZhang et al. (2025) | .916 | **.903** | - | **.975** | **.016** | **.889** | **.870** | - | **.944** | **.036** | **.893** | **.849** | - | **.953** | **.015** | **.903** | **.883** | - | **.951** | **.023** |
| ***Semi-Supervised Methods*** | | | | | | | | | | | | | | | | | | | | |
| CamoTeacher(10%)Lai et al. (2024) | .756 | .617 | .684 | .813 | .065 | .701 | .560 | .742 | .795 | .112 | .759 | .594 | .836 | .854 | .049 | .791 | .687 | .842 | .868 | .068 |
| SCOD-ND(10%)Fu et al. (2024) | .850 | .773 | - | .928 | .036 | .789 | .732 | - | .859 | .077 | .819 | .725 | - | .891 | .033 | .838 | .787 | - | .903 | .046 |
| SCOUT(10%)Yan et al. (2025b) | .874 | .815 | .838 | .916 | .027 | .847 | .812 | .849 | .901 | .052 | .858 | .783 | .815 | .917 | .024 | .873 | .833 | .864 | .919 | .035 |
| ***Unsupervised Methods*** | | | | | | | | | | | | | | | | | | | | |
| BigGW Voynov et al. (2021) | .547 | .244 | .294 | .527 | .257 | .565 | .299 | .349 | .528 | .282 | .528 | .185 | .246 | .497 | .261 | .608 | .319 | .391 | .565 | .246 |
| TokenCutWang et al. (2022) | .654 | .496 | .536 | .740 | .132 | .633 | .498 | .543 | .706 | .163 | .658 | .469 | .502 | .735 | .103 | .725 | .615 | .649 | .802 | .101 |
| TokenCut w/B.S.Wang et al. (2022) | .655 | .351 | .393 | .582 | .169 | .639 | .383 | .434 | .595 | .195 | .666 | .334 | .399 | .609 | .127 | .735 | .478 | .547 | .683 | .133 |
| SpectralSegMelas-Kyriazi et al. (2022) | .575 | .410 | .440 | .628 | .220 | .579 | .450 | .481 | .648 | .235 | .575 | .360 | .388 | .595 | .193 | .669 | .535 | .562 | .719 | .159 |
| SelfMaskShin et al. (2022) | .619 | .436 | .481 | .675 | .176 | .617 | .483 | .536 | .698 | .176 | .637 | .431 | .469 | .679 | .131 | .716 | .593 | .634 | .777 | .114 |
| SelfMask w/U.B.Shin et al. (2022) | .629 | .447 | .491 | .683 | .169 | .627 | .495 | .547 | .708 | .182 | .645 | .440 | .478 | .687 | .125 | .723 | .601 | .642 | .784 | .110 |
| *FOUNDSiméoni et al. (2023) | .829 | .757 | .781 | .911 | .040 | .770 | .704 | .740 | .849 | .090 | .767 | .641 | .668 | .847 | .045 | .816 | .756 | .783 | .893 | .052 |
| *UCOS-DAZhang & Wu (2023) | .750 | .639 | .666 | .808 | .091 | .702 | .604 | .633 | .751 | .148 | .655 | .467 | .495 | .687 | .120 | .731 | .617 | .644 | .785 | .103 |
| RISEDu et al. (2025b) | .805 | .675 | - | .858 | .052 | .722 | .587 | - | .775 | .113 | .753 | .578 | - | .827 | .053 | .797 | .687 | - | .860 | .064 |
| EASEDu et al. (2025a) | .819 | .741 | - | .899 | .044 | .749 | .684 | - | .831 | .098 | .773 | .656 | - | .866 | .040 | .800 | .735 | - | .884 | .056 |
| UCOD-DPLYan et al. (2025a) | .864 | .825 | .838 | **.931** | .031 | .793 | .747 | .779 | **.862** | .077 | .834 | .763 | .779 | .916 | .031 | .850 | .818 | .835 | .923 | .043 |
| **Ours** | **.877** | **.831** | **.848** | **.931** | **.026** | **.812** | **.762** | **.802** | .858 | **.071** | **.859** | **.798** | **.823** | **.934** | **.023** | **.876** | **.848** | **.870** | **.935** | **.033** |

**Bold** indicates the best result in group settings, and underline indicates the second-best result. * denotes the version that is reimplemented by us.

Based on the skeleton-induced detail exploration regions $\Omega$, we apply an adaptive local temperature rescaling mechanism to refine fine-grained details within each component $C_k \in \Omega$.

$$\begin{aligned} \phi_T(p;\mu_k) &= \sigma\left(\frac{\mathrm{logit}(p)-\mathrm{logit}(\mu_k)}{T}\right), \mu_k = \frac{1}{|C_k|}\sum_{(x,y)\in C_k}\hat{P}_i^s(x,y), \\ \hat{P}_i^{TR}(x,y) &= \begin{cases} \phi_T\left(\hat{P}_i^s(x,y);\mu_k\right), & (x,y)\in\Omega, \\ \hat{P}_i^s(x,y), & (x,y)\notin\Omega. \end{cases} \end{aligned} \tag{12}$$

where $\mathrm{logit}(\cdot)$ represents the inverse sigmoid. $\odot$ denotes element-wise multiplication, and $T<1$ controls the sharpening strength. Pixels outside $\Omega$ remain unchanged.

Finally, $\hat{P}_i^{TR}$ is used as a fine-grained pseudo-label to supervise the student model:

$$\mathcal{L}_{ADE} = \mathcal{L}_{Seg}\left(\hat{P}_i^s(x,y), \hat{P}_i^{TR}(x,y)\right), \tag{13}$$

### 3.5 Total loss

Finally, the total loss $\mathcal{L}_{tot}$ can be defined as follows:

$$\mathcal{L}_{\mathrm{tot}} = \mathcal{L}_{\mathrm{STI}} + \mathcal{L}_{\mathrm{SCD}} + \mathcal{L}_{\mathrm{ADE}}, \tag{14}$$

where $\mathcal{L}_{\mathrm{STI}}$ in eq. (4) is to optimize the teacher initialization. $\mathcal{L}_{\mathrm{SCD}}$ in eq. (9) enforces subject consistency and suppresses confusion, and $\mathcal{L}_{\mathrm{ADE}}$ in eq. (13) refines fine structures via detail exploration.

Following Zheng et al. (2024), we use a combination of Binary Cross Entropy loss $\mathcal{L}_{\mathrm{BCE}}$, Intersection Over Union loss $\mathcal{L}_{\mathrm{IOU}}$, and Structure Similarity Index Measure loss $\mathcal{L}_{\mathrm{SSIM}}$ to build the segmentation loss $\mathcal{L}_{\mathrm{Seg}}$ in eq. (7).

During the subject localization with distillation phase, we first initialize the teacher model using $\mathcal{L}_{\mathrm{STI}}$. Then we keep the teacher frozen and optimize the student model with $\mathcal{L}_{\mathrm{SCD}}$. During the detail exploration phase, we train the student model with $\mathcal{L}_{\mathrm{ADE}}$.

## 4 Experiments

### 4.1 Experimental Settings

**Datasets and Evaluation Protocol.** Following Luo et al. (2023); Fan et al. (2020a), we train on 1,000 CAMO and 3,040 COD10K training images without ground-truth annotations. Evaluation is

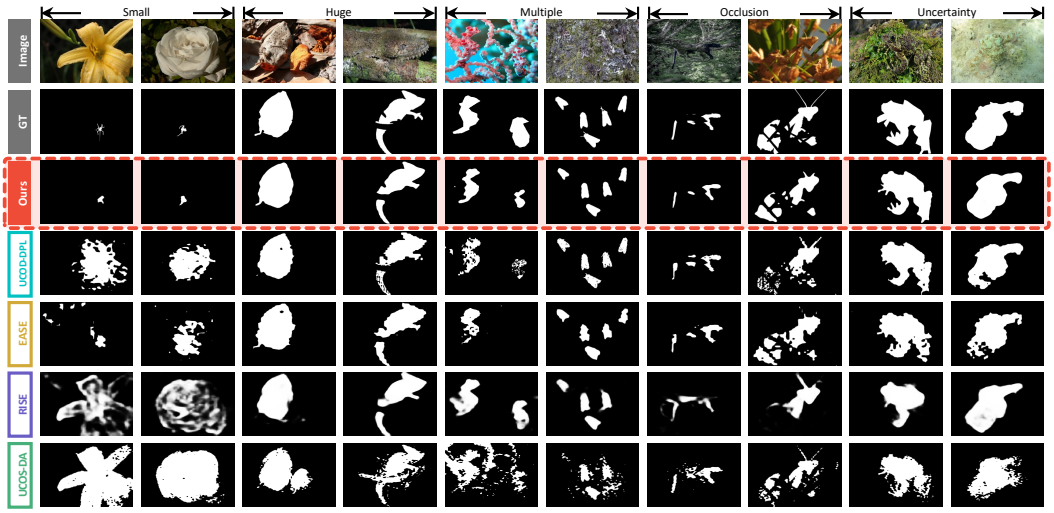



Figure 4: Visual comparison of our method with other existing methods in challenging scenarios. Our method (third row, highlighted with a red background) achieves higher segmentation accuracy and produces clearer object edges and detailed texture compared to existing approaches.

conducted on four COD benchmarks: CHAMELEON (76 images), CAMO-Test (250), COD10K-Test (2,026), and NC4K (4,121). We report S-measure ($\mathcal{S}_m$) Fan et al. (2017), mean and weighted F-measure ($\mathcal{F}_\beta^m$, $\mathcal{F}_\beta^\omega$) Margolin et al. (2014), mean E-measure ($\mathcal{E}_\epsilon^m$) Fan et al. (2018), and mean absolute error ($\mathcal{M}$) Perazzi et al. (2012).

**Implementation Details.** We use pretrained DINOv2 Oquab et al. (2023) as the backbone, with minor adaptations to its intermediate layers for hierarchical feature extraction. The teacher uses DBA Yan et al. (2025a) with last-layer features, while the student uses BiRefBlock Zheng et al. (2024) with hierarchical features. AdamW with $eps = 1e-8$ is used as the optimizer. The learning rate is set to 2e-4 for STI, 1e-4 for SCD and ADE. We conduct two training rounds, where the initial round uses pseudo-labels from a fixed strategy and the following round uses pseudo-labels from the student predictions of the previous round. The STI is trained for 25 epochs per round, and the SCD for 6 epochs. The batch size is set to 4 per GPU. All experiments are implemented in PyTorch 2.1 and executed on a workstation equipped with an Intel(R) Xeon(R) Silver 4214R CPU and four NVIDIA A800 GPUs. More implementation details will be provided in the Appendix A.

### 4.2 Main Results

**Qualitative Analysis.** We show visualization comparisons of our model's predictions and those of other competing methods across a range of challenging scenarios. As illustrated in fig. 4, our approach demonstrates finer boundary and structural details, more coherent segmentation within objects, and higher segmentation quality.

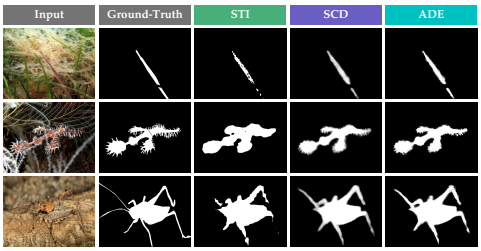



Figure 5: **Visualization of predictions in different modules.**

**Quantitative Analysis.** We quantitatively compare our proposed SEED with state-of-the-art (SOTA) methods in Unsupervised Semantic Segmentation (USS) and Unsupervised Camouflaged Object Detection (UCOD) across four COD datasets. To further highlight the competitiveness of our approach, we also include comparisons against existing semi-supervised and fully-supervised methods. As shown in table 1, the proposed SEED consistently outperforms existing UCOD methods across most evaluation metrics and datasets, achieving SOTA performance. Notably, on the two large-scale benchmarks, i.e., COD10K-Test and NC4K, SEED achieves clear and consistent improvements over UCOD-DPL across all major metrics. Compared with UCOD-DPL, SEED achieves an average gain of 2.47% in $\mathcal{S}_m$ and a 23.02% reduction in MAE. This consistent superiority across diverse datasets demonstrates SEED's robust generalization capability in handling diverse camouflage scenarios.

### 4.3 Ablation Study

To verify the effectiveness of our methods, we conduct comprehensive ablation studies on the COD10K-Test dataset.

Table 2: Ablation study to evaluate the proposed modules.

| | Settings | | | COD10K (2026) | | | | |
|---|---|---|---|---|---|---|---|---|
| | STI | SCD | ADE | $\mathcal{S}_m \uparrow$ | $\mathcal{F}_\beta^\omega \uparrow$ | $\mathcal{F}_\beta^m \uparrow$ | $\mathcal{E}_\phi^m \uparrow$ | $\mathcal{M} \downarrow$ |
| **1st-Round** | | | | .498 | .290 | .320 | .500 | .261 |
| | ✓ | | | .837 | .758 | .775 | .912 | .028 |
| | ✓ | ✓ | | .842 | .765 | .784 | .913 | .028 |
| | ✓ | | ✓ | .837 | .758 | .775 | .913 | .028 |
| | ✓ | ✓ | ✓ | **.850** | **.779** | **.797** | **.927** | **.026** |
| **2nd-Round** | ✓ | | | .840 | .766 | .790 | .914 | .027 |
| | ✓ | ✓ | | .843 | .771 | .797 | .914 | .026 |
| | ✓ | | ✓ | .841 | .767 | .789 | .919 | .027 |
| | ✓ | ✓ | ✓ | **.859** | **.798** | **.823** | **.934** | **.023** |

**Module Ablation.** We conduct module ablations in table 2 between STI, SCD and ADE. Specifically, for STI ablation, we replace the teacher model with fixed-strategy pseudo-labels, and for SCD ablation, we directly use the segmentation loss between the teacher and student. Without the STI, the initial pseudo-labels derived from pre-trained features are heavily contaminated by background noise, which severely hinders the student model from accurately localizing the camouflaged subject. By employing a dynamic filtering strategy, the STI module effectively mitigates these distractors. Furthermore, we observe a strong interdependence between SCD and ADE modules. Specifically, SCD tends to produce over-smoothed boundaries due to its emphasis on consistency; conversely, while ADE sharpens details via temperature rescaling, it risks amplifying noise if applied to impure features, erroneously enhancing background textures as object edges. Therefore, combining these two modules yields a substantial performance leap. As shown in fig. 5, the visualization further verifies the role of each module: STI suppresses background distractors, SCD improves subject consistency, and ADE enhances fine details. In addition, the second training round outperforms the first as the model trained in the first round generates more accurate predictions, providing higher-quality supervision for further optimization.

Table 3: Ablation study on different backbone.

| Methods | COD10K (2026) | | | | |
|---|---|---|---|---|---|
| | $\mathcal{S}_m \uparrow$ | $\mathcal{F}_\beta^\omega \uparrow$ | $\mathcal{F}_\beta^m \uparrow$ | $\mathcal{E}_\phi^m \uparrow$ | $\mathcal{M} \downarrow$ |
| UCOS-DA$_{DINOv1}$ | .689 | .513 | .546 | .740 | .086 |
| UCOD-DPL$_{DINOv1}$ | .727 | .577 | .627 | **.822** | .059 |
| **Ours**$_{DINOv1}$ | **.729** | **.579** | **.646** | .815 | **.052** |
| UCOS-DA$_{DINOv2}$ | .655 | .467 | .495 | .687 | .120 |
| UCOD-DPL$_{DINOv2}$ | .834 | .763 | .779 | .916 | .031 |
| **Ours**$_{DINOv2}$ | **.859** | **.798** | **.823** | **.934** | **.023** |

Table 4: Ablation study on BIoU.

| Methods | CHAMELEON | CAMO | COD10K | NC4K |
|---|---|---|---|---|
| EASE | .401 | .307 | .380 | .388 |
| RISE | .425 | .317 | .374 | .403 |
| UCOD-DPL | .530 | .398 | .535 | .503 |
| UCOD-DPL$_{\text{woLT}}$ | .462 | .365 | .410 | .434 |
| **SEED** | **.576** | **.468** | **.591** | **.587** |

**woLT** denotes the variant without using the $Look-Twice$ strategy.

**Study on Different Backbone.** We conduct a backbone replacement study on the different backbones for UCOS-DA Zhang & Wu (2023), UCOD-DPL Yan et al. (2025a), and the SEED in table 3. The SEED demonstrates advantages over other UCOD methods under different settings.

**Boundary Quality Assessment** As shown in table 4, SEED not only reduces background noise across region-based metrics (e.g., MAE) but also significantly improves the boundary metric (BIoU) Cheng et al. (2021), proving that our decoupling strategy effectively breaks the inherent trade-off.

## 5 Conclusion

In this paper, we address the inherent trade-off between noise suppression and detail preservation in existing UCOD methods: conservative filtering fails to remove noise that disrupts subject localization, whereas aggressive noise filtering achieves subject purification but sacrifices critical edge and structural details. Therefore, we decompose UCOD into subject localization with distillation phase and detail exploration phase. Accordingly, we propose a Teacher–Student framework for Subject-to-Edge Exploration and Distillation with Subject Teacher Initialization (STI) module, Subject Consistency Distillation (SCD) module and Active Detail Exploration (ADE) module. During subject localization with distillation phase, the STI employs a dynamic pseudo-label to obtain purified camouflaged subject. Then, SCD enforces teacher–student consistency in high-confidence regions and uses confusion suppression in confusion regions to stabilize the camouflaged subject and mitigate background noise. During the detail exploration phase, ADE uses adaptive local temperature rescaling to refine boundary predictions. Experiments show that SEED outperforms existing UCOD methods and is competitive with several fully supervised approaches.

### AI USE STATEMENT

In this work, we used generative AI tools to design and provide feedback on experiments, assist with method implementation, and assist with translation. We did not use generative AI tools for any other tasks requiring disclosure under the ICLR 2027 AI Policy, and required-disclosure tasks not involved in this work were not applicable. Additionally, we used generative AI tools to edit software code, draft portions of the manuscript, and improve the readability of the paper.

All AI-generated and AI-assisted content was reviewed and evaluated by at least two authors. The authors assessed the correctness, appropriateness, and consistency of the AI-assisted outputs with the implemented methods and experimental evidence, and revised or discarded them where necessary. We take full responsibility for the final content of this work, including its text, claims, code, methodology, experimental design, results, and other artifacts produced with the aid of generative AI.

## REFERENCES

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Alan B. Cannaday, Curt H. Davis, and Trevor M. Bajkowski. Detection of camouflage-covered military objects using high-resolution multi-spectral satellite imagery. In *IGARSS 2023*, pp. 5766–5769, Pasadena, CA, USA, 2023. IEEE. ISBN 979-8-3503-3174-5. doi: 10.1109/IGARSS52108.2023.10281409.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Int. Conf. Comput. Vis.*, pp. 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15334–15342, 2021.

Ji Du, Fangwei Hao, Mingyang Yu, Desheng Kong, Jiesheng Wu, Bin Wang, Jing Xu, and Ping Li. Shift the lens: Environment-aware unsupervised camouflaged object detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19271–19282, Nashville, TN, USA, 2025a. IEEE. ISBN 979-8-3315-4365-5. doi: 10.1109/CVPR52734.2025.01795.

Ji Du, Xin Wang, Fangwei Hao, Mingyang Yu, Chunyuan Chen, Jiesheng Wu, Bin Wang, Jing Xu, and Ping Li. Beyond single images: Retrieval self-augmented unsupervised camouflaged object detection. *CoRR*, abs/2510.18437, 2025b. doi: 10.48550/ARXIV.2510.18437.

Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Int. Conf. Comput. Vis.*, pp. 4548–4557, 2017.

Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 698–704. AAAI Press, 2018.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2774–2784, 2020a. doi: 10.1109/CVPR42600.2020.00285.

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 *International conference on medical image computing and computer-assisted intervention*, pp. 263–273. Springer, 2020b.

Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042, 2022. doi: 10.1109/TPAMI.2021.3085766.

Yuanbin Fu, Jie Ying, Houlei Lv, and Xiaojie Guo. Semi-supervised camouflaged object detection from noisy data. In *ACM Int. Conf. Multimedia*, 2024.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 16000–16009, 2022.

Xiaobin Hu, Shuo Wang, Xuebin Qin, Hang Dai, Wenqi Ren, Donghao Luo, Ying Tai, and Ling Shao. High-resolution iterative feedback network for camouflaged object detection. In *AAAI*, volume 37, pp. 881–889, 2023.

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 Conf. Comput. Vis. Pattern Recog.*, pp. 5557–5566, 2023.

Shuo Jiang, Gaojia Zhang, Min Tan, Yufei Yin, and Gang Pan. Erecu: Pseudo-label evolution fusion and refinement with multi-cue learning for unsupervised camouflage detection. *arXiv preprint arXiv:2603.11521*, 2026.

Xunfa Lai, Zhiyu Yang, Jie Hu, Shengchuan Zhang, Liujuan Cao, Guannan Jiang, Zhiyu Wang, Songan Zhang, and Rongrong Ji. Camoteacher: Dual-rotation consistency learning for semi-supervised camouflaged object detection. In *European Conference on Computer Vision*, 2024.

Bing Li, Rongqian Zhou, Lu Yang, Qiwen Wang, and Huang Chen. Mildetr: Detection transformer for military camouflaged target detection. *IEEE Access*, 12:26163–26174, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3363442.

Shijie Liu, Binwei Xu, Tuo Shen, Guanghui Yue, and Qiuping Jiang. Self-anchored progressive framework with noise mitigation for unsupervised camouflaged object detection. *IEEE Transactions on Image Processing*, 35:3581–3593, 2026.

Yi Liu, Chengxin Li, Xiaohui Dong, Lei Li, Dingwen Zhang, Shoukun Xu, and Jungong Han. Seamless detection: Unifying salient object detection and camouflaged object detection. *CoRR*, abs/2412.16840, 2024. doi: 10.48550/ARXIV.2412.16840.

Naisong Luo, Yuwen Pan, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Camouflaged instance segmentation via explicit de-camouflaging. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 17918–17927, 2023.

Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 248–255, 2014.

Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8364–8375, 2022.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 733–740. IEEE, 2012.

Gyungin Shin, Samuel Albanie, and Weidi Xie. Unsupervised salient object detection with spectral cluster voting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3971–3980, 2022.

Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. cldice - A novel topology-preserving loss function for tubular structure segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 16560–16569. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01629.

Peiyao Shou, Yixiu Liu, Wei Wang, Yaoqi Sun, Zhigao Zheng, Shangdong Zhu, and Chenggang Yan. Sdalsnet: Self-distilled attention localization and shift network for unsupervised camouflaged object detection. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 6914–6921. AAAI Press, 2025. doi: 10.1609/AAAI.V39I7.32742.

Oriane Siméoni, Chloé Sekkat, Gilles Puy, Antonín Vobeckỳ, Éloi Zablocki, and Patrick Pérez. Unsupervised object localization: Observing the background to discover objects. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3176–3186, 2023.

Yujia Sun, Shuo Wang, Chenglizhao Chen, and Tian-Zhu Xiang. Boundary-guided camouflaged object detection. In Lud De Raedt (ed.), *The International Joint Conference on Artificial Intelligence*, pp. 1335–1341. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/186. Main Track.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

Andrey Voynov, Stanislav Morozov, and Artem Babenko. Object segmentation without labels with large-scale generative models. In *International Conference on Machine Learning*, pp. 10596–10606. PMLR, 2021.

Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 14543–14553, 2022.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9653–9663, 2022.

Weiqi Yan, Lvhai Chen, Huaijia Kou, Shengchuan Zhang, Yan Zhang, and Liujuan Cao. Ucod-dpl: Unsupervised camouflaged object detection via dynamic pseudo-label learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30365–30375, 2025a.

Weiqi Yan, Lvhai Chen, Shengchuan Zhang, Yan Zhang, and Liujuan Cao. Scout: Semi-supervised camouflaged object detection by utilizing text and adaptive data selection. In *International Joint Conference on Artificial Intelligence*, 2025b.

Shizhou Zhang, Dexuan Kong, Yinghui Xing, Yue Lu, Lingyan Ran, Guoqiang Liang, Hexu Wang, and Yanning Zhang. Frequency-guided spatial adaptation for camouflaged object detection. *IEEE Transactions on Multimedia*, 27:72–83, 2025. ISSN 1941-0077. doi: 10.1109/TMM.2024.3521681.

T. Y. Zhang and C. Y. Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3):236–239, 1984.

Yi Zhang and Chengyi Wu. Unsupervised camouflaged object segmentation as domain adaptation. In *Int. Conf. Comput. Vis.*, pp. 4334–4344, 2023.

Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3:9150038, 2024. doi: 10.26599/AIR.2024.9150038.

## A Implementation Details.

All experiments are implemented with PyTorch 2.1, and run on a machine with Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz, 512GiB RAM, and four NVIDIA Titan A800-80G GPUs. All experiments use the same random seed. We set $\tau_{ce} = 0.7$ and $K = 0.2$ as fixed, non-tuned default hyperparameters for all experiments. We use AdamW with $\epsilon = 1e-8$ as the optimizer. The learning rate is set to $2e-4$ for STI and $1e-4$ for SCD and ADE. We conduct two training rounds: in the first round, pseudo-labels are generated using a fixed strategy, while in the second round, pseudo-labels are obtained from the student predictions produced in the previous round. The STI model is trained for 25 epochs per round, and the SCD model is trained for 6 epochs per round.

We build the network by using DINOv2-Base Oquab et al. (2023) as the backbone to extract image features, which are then processed by subsequent modules. All backbone parameters remain frozen during the training process. The teacher model adopts a lightweight DBA decoder proposed from UCOD-DPL Yan et al. (2025a) and uses only the last-layer features from DINOv2. In contrast, the student model employs the more complex BiRefBlock proposed from BiRefNet Zheng et al. (2024) and leverages the hierarchical multi-scale features of DINOv2. To obtain hierarchical multi-scale features from DINOv2, we extract features from the 2nd, 5th, 8th, and 11th layers of DINOv2, where the layer indices are counted from 0. These features are then transformed in terms of spatial resolution and channel dimension using simple bilinear interpolation and $1 \times 1$ convolutions, respectively, with the $1 \times 1$ convolution parameters learnable during training. As a result, we obtain multi-level feature representations with resolutions of $[H/2, H/4, H/8, H/16]$ and channel dimensions of $[C/4, C/2, C, C]$, where $H$ denotes the input image resolution, and $C$ denotes the channel dimension of the last-layer features from DINOv2.

## B More Experiments

**Hyperparameter Sensitivity Analysis.** We investigate the sensitivity of our method to the certainty threshold $\tau_{\text{ce}}$ in section 3.3. As shown in fig. 6, the model maintains stable performance across a range of $\tau_{\text{ce}}$ values, indicating that our method is not sensitive to the choice of $\tau_{\text{ce}}$.

We further study the sensitivity to different $K$ values in Top-/Bottom-$K$ selection, as shown in fig. 7. The values used for Top-$K$ and Bottom-$K$ are selected symmetrically. The results show that our method is relatively insensitive to small $K$ values. This suggests that the Top-/Bottom-$K$ design can serve as a warm-up strategy, guiding the student toward confident predictions in ambiguous regions during the early training stage. However, an excessively large $K$ may force the model to push noisy pixels toward confidence extremes, leading to performance degradation.

The sensitivity analysis in the supplementary material is used only to verify robustness, showing that SEED maintains stable performance within a reasonable range of these hyperparameters.

| $\tau_{ce}$ | COD10K (2026) | | | | |
|---|---|---|---|---|---|
| | $\mathcal{S}_m \uparrow$ | $\mathcal{F}_\beta^\omega \uparrow$ | $\mathcal{F}_\beta^m \uparrow$ | $\mathcal{E}_\phi^m \uparrow$ | $\mathcal{M} \downarrow$ |
| 0.6 | .856 | .794 | .813 | .931 | .024 |
| 0.65 | .858 | .797 | .822 | .929 | .023 |
| **0.7** | .859 | .798 | .823 | .934 | .023 |
| 0.75 | .855 | .795 | .818 | .930 | .024 |
| 0.8 | .857 | .798 | .818 | .927 | .024 |

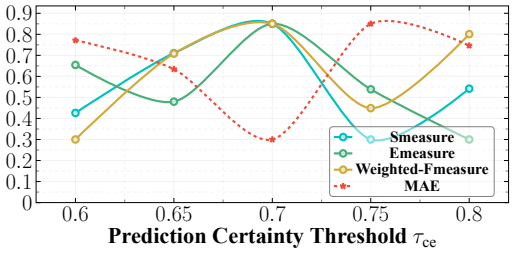



Figure 6: **Hyperparameter sensitivity analysis of the certainty threshold $\tau_{\text{ce}}$.** The values of each metric are normalized only for visualization, enabling all metrics to be compared in a single figure.

**Analysis of Efficiency.** SEED deliberately adopts an asymmetric teacher-student architecture, which is not a trivial scaling-up design but a key component of our decoupled localization-and-detail learning strategy. Specifically, the teacher is responsible for robust subject localization. Therefore, it employs a lightweight decoder, i.e., DBA, and mainly relies on high-level semantic features to form an information bottleneck. This design helps suppress low-level background noise and prevents the model from overfitting to inaccurate pseudo-label details during the early training stage. In contrast,

| $K$ | COD10K (2026) | | | | |
|---|---|---|---|---|---|
| | $\mathcal{S}_m \uparrow$ | $\mathcal{F}_\beta^\omega \uparrow$ | $\mathcal{F}_\beta^m \uparrow$ | $\mathcal{E}_\phi^m \uparrow$ | $\mathcal{M} \downarrow$ |
| 0.1 | .862 | .795 | .822 | .931 | .024 |
| **0.2** | .859 | .798 | .823 | .934 | .023 |
| 0.3 | .853 | .783 | .815 | .923 | .025 |
| 0.4 | .848 | .781 | .810 | .921 | .026 |

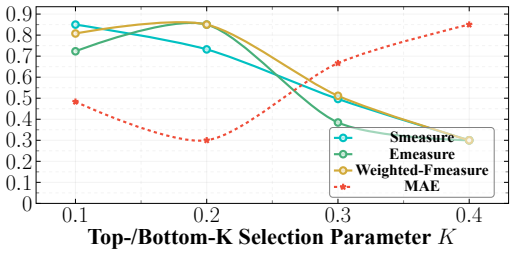



Figure 7: **Hyperparameter sensitivity analysis of the Top-/Bottom-$K$ selection parameter $K$.** The values of each metric are normalized only for visualization, enabling all metrics to be compared in a single figure.

the student is designed for detail exploration. Guided by the teacher's subject-level localization, the student progressively recovers fine-grained boundaries and subtle structural details, which requires stronger representation capacity and hierarchical multi-scale feature aggregation. Therefore, the larger student decoder based on BiRefBlock is a purposeful design to resolve the noise-detail trade-off, rather than an unfair increase in model capacity.

During inference, all auxiliary components are discarded, and only the student model, including the encoder and decoder, is retained for end-to-end prediction. Thus, the evaluated cost of SEED(Student) accurately reflects the real deployment cost. As shown in table 5, SEED(Student) achieves 10.6 FPS, indicating that the proposed method still maintains real-time inference efficiency despite using a stronger decoder for detail recovery. Moreover, compared with state-of-the-art fully supervised and semi-supervised methods, where the latter use 10% labeled data, SEED(Student) requires substantially fewer trainable parameters while outperforming current semi-supervised leaders. These results demonstrate that SEED achieves a favorable balance between model efficiency and segmentation accuracy.

It is also worth noting that SEED(Teacher) adopts the same model architecture as UCOD-DPL. Nevertheless, our method still achieves superior performance, which further verifies that the improvement does not simply come from architectural scaling, but from the proposed decoupled localization-and-detail learning framework.

Table 5: Comparison of model scale and computational overhead.

| Methods | Backbone | $S_m \uparrow$ | $\mathcal{M} \downarrow$ | Trainable Params (Total Params) | FPS | MACs |
|---|---|---|---|---|---|---|
| **Fully-Supervised Methods** | | | | | | |
| BiRefNet | Swin-B | 0.913 | 0.014 | 106.23M (106.23M) | 11.3 | 559.99G |
| **Semi-Supervised Methods** | | | | | | |
| SCOUT (10%) | Swin-B | 0.858 | 0.024 | 107.50M (411.01M) | 10.4 | 386.87G |
| **Unsupervised Methods** | | | | | | |
| UCOS-DA | DINOv2-B | 0.655 | 0.120 | 2.94M (89.52M) | 27.2 | 111.2G |
| UCOD-DPL | DINOv2-B | 0.834 | 0.031 | 0.10M (86.59M) | 17.6 | 117.5G |
| **SEED(Teacher)** | DINOv2-B | 0.844 | 0.027 | 0.10M (86.95M) | 17.6 | 117.5G |
| **SEED(Student)** | DINOv2-B | 0.859 | 0.023 | 14.74M (101.32M) | 10.6 | 425.59G |

**Study on Multi-round Training.** We further investigate the effect of multi-round training by evaluating both the teacher and student models after each training round on four COD benchmarks. As shown in table 6, the performance improves substantially from the 1st round to the 2nd round, indicating that the predictions generated after the first round provide higher-quality supervision for subsequent optimization. However, continuing to the 3rd round brings only marginal and inconsistent changes: although a few metrics slightly improve, the student performance degrades on CAMO-Test and COD10K-Test in terms of several key metrics, including $\mathcal{S}_m$, $E_m^\phi$, and MAE. This suggests that the training process has largely converged after the 2nd round, and further rounds may introduce accumulated pseudo-label noise or over-refinement.

Table 6: Ablation on multi-round training.

| Training Round | Model | CHAMELEON | | | | | CAMO-Test | | | | | COD10K-Test | | | | | NC4K | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\mathcal{S}_m \uparrow$ | $\mathcal{F}_\beta^\omega \uparrow$ | $\mathcal{F}_\beta^m \uparrow$ | $\mathcal{E}_\phi^m \uparrow$ | $\mathcal{M} \downarrow$ | $\mathcal{S}_m \uparrow$ | $\mathcal{F}_\beta^\omega \uparrow$ | $\mathcal{F}_\beta^m \uparrow$ | $\mathcal{E}_\phi^m \uparrow$ | $\mathcal{M} \downarrow$ | $\mathcal{S}_m \uparrow$ | $\mathcal{F}_\beta^\omega \uparrow$ | $\mathcal{F}_\beta^m \uparrow$ | $\mathcal{E}_\phi^m \uparrow$ | $\mathcal{M} \downarrow$ | $\mathcal{S}_m \uparrow$ | $\mathcal{F}_\beta^\omega \uparrow$ | $\mathcal{F}_\beta^m \uparrow$ | $\mathcal{E}_\phi^m \uparrow$ | $\mathcal{M} \downarrow$ |
| **1st-Round** | Teacher | .841 | .775 | .786 | .905 | .034 | .791 | .737 | .767 | .865 | .076 | .784 | .675 | .692 | .869 | .036 | .827 | .776 | .792 | .904 | .045 |
| | Student | .870 | .814 | .828 | .923 | .029 | .809 | .744 | .781 | .857 | .076 | .850 | .779 | .797 | .927 | .026 | .872 | .829 | .847 | .928 | .036 |
| **2nd-Round** | Teacher | .867 | .833 | .848 | .930 | .030 | .800 | .760 | .794 | .864 | .073 | .844 | .780 | .799 | .921 | .027 | .858 | .831 | .851 | .926 | .038 |
| | Student | .877 | .831 | .848 | .931 | .026 | .812 | .762 | .802 | .858 | .071 | .859 | .798 | .823 | .934 | .023 | .876 | .848 | .870 | .935 | .033 |
| **3rd-Round** | Teacher | .868 | .836 | .857 | .930 | .030 | .794 | .756 | .795 | .856 | .074 | .845 | .785 | .808 | .921 | .026 | .857 | .832 | .857 | .923 | .039 |
| | Student | .881 | .826 | .853 | .938 | .027 | .807 | .742 | .789 | .838 | .076 | .856 | .798 | .832 | .926 | .024 | .876 | .838 | .871 | .923 | .035 |

**Pseudo-Label Quality Across Training Progress.** To analyze the evolution of pseudo-label quality during training, we evaluate the ROI prediction $\hat{P}_i^{\mathrm{ROI}}$ defined in eq. (6) and the final supervision target $T_i$ defined in eq. (7) against the ground-truth masks at five training checkpoints. The ground-truth masks are used only for this analysis and are never involved in model training.

As shown in table 7, the quality of both $\hat{P}_i^{\mathrm{ROI}}$ and $T_i$ improves steadily throughout training. At the early stage, $T_i$ exhibits more reliable region-level quality, achieving a substantially lower MAE than $\hat{P}_i^{\mathrm{ROI}}$. Meanwhile, the boundary quality of $\hat{P}_i^{\mathrm{ROI}}$ progressively improves, and its performance gradually converges with that of $T_i$. At the end of training, the two signals achieve nearly identical region- and boundary-level performance. The consistent improvements in both $\mathcal{S}_m$ and BIoU demonstrate stable pseudo-label refinement without evident error accumulation.

Table 7: Pseudo-label quality and training stability at different training stages on the combined TR-COD10K and TR-CAMO training sets.

| Progress | ROI Prediction $\hat{P}^{\mathrm{ROI}}$ | | | Supervision Target $T$ | | |
|---|---|---|---|---|---|---|
| | $\mathcal{S}_m \uparrow$ | $\mathcal{M} \downarrow$ | BIoU $\uparrow$ | $\mathcal{S}_m \uparrow$ | $\mathcal{M} \downarrow$ | BIoU $\uparrow$ |
| 20% | .815 | .069 | .521 | .847 | .035 | .518 |
| 40% | .867 | .035 | .583 | .858 | .034 | .522 |
| 60% | .873 | .033 | .589 | .868 | .033 | .554 |
| 80% | .875 | **.033** | .592 | .874 | .033 | .589 |
| 100% | **.875** | **.033** | **.594** | **.876** | **.033** | **.595** |

**Ablation Study of the Asymmetric Teacher–Student Architecture.** We further investigate the necessity of the asymmetric teacher–student architecture on COD10K-Test dataset. The teacher is trained with relatively low-quality pseudo-labels that contain substantial noise. It therefore adopts a DBA decoder that uses only the last-layer feature $F_4$. As a high-level feature, $F_4$ contains strong semantic information and is less affected by the noisy details in the pseudo-labels, making it suitable for stable subject localization. In contrast, the multi-level BiRefBlock uses features from different levels and preserves more spatial information. Although these spatial details are useful for refined prediction, they are also more easily affected by the noise in the pseudo-labels during teacher training. Replacing the teacher's DBA decoder with the multi-level BiRefBlock therefore causes a substantial performance drop, reducing $S_\alpha$ from .784 to .449.

For the student, using the same DBA decoder achieves an $S_\alpha$ of .842, showing that the performance gain mainly comes from SEED's training strategy rather than from the decoder design alone. Replacing the student's DBA decoder with the multi-level BiRefBlock further improves the results to $S_\alpha = .850$, $F_\beta^w = .779$, $F_\beta^{\max} = .797$, and $E_\phi^{\max} = .927$, while reducing MAE to .026. These results demonstrate the benefit of the asymmetric teacher–student framework: the teacher uses high-level semantic features for stable subject localization, while the student uses multi-level features to recover detailed object structures after the subject prediction has been refined.

Table 8: Ablation study of the asymmetric teacher–student architecture on COD10K-Test.

| Model Role | Decoder | Input Features | $\mathcal{S}_m \uparrow$ | $\mathcal{F}_\beta^\omega \uparrow$ | $\mathcal{F}_\beta^m \uparrow$ | $\mathcal{E}_\phi^m \uparrow$ | $\mathcal{M} \downarrow$ |
|---|---|---|---|---|---|---|---|
| Teacher | BiRefBlock | Hierarchical | .449 | .199 | .234 | .449 | .272 |
| Teacher | DBA | Last-layer | .784 | .675 | .692 | .869 | .036 |
| Student | DBA | Last-layer | .842 | .741 | .789 | .921 | .028 |
| Student | BiRefBlock | Hierarchical | **.850** | **.779** | **.797** | **.927** | **.026** |

**Ablation Study of Feature Selection in STI.** STI is designed to establish a reliable subject prior rather than recover complete object details at the teacher stage. This motivates the use of different feature configurations for the teacher and the student. To examine this design, we compare the last-layer feature and multi-level features under fully supervised and unsupervised settings. For the multi-level DBA variant, we project the multi-level features to the same channel dimension, fuse them by element-wise addition, and feed the fused feature into the unchanged DBA decoder. BiRefBlock uses its original hierarchical fusion scheme. All other settings remain unchanged.

As shown in Table 9, multi-level features are effective when clean ground-truth supervision is available. DBA with multi-level features achieves $S_\alpha = .843$, compared with .841 for DBA with the last-layer feature, while BiRefBlock with multi-level features reaches .869. This result is consistent with the benefit of multi-level features for supervised segmentation, where additional spatial information can be learned from reliable labels.

However, the same feature choice is not suitable for teacher training in UCOD. The initial pseudo-labels contain substantial noise, with $S_\alpha = .483$ and MAE $= .295$. Under unsupervised training, DBA with the last-layer feature achieves $S_\alpha = .784$, whereas DBA with multi-level features drops to .607. BiRefBlock with multi-level features further drops to .449. This contrast shows that the role of multi-level features depends on the quality of supervision. Lower-level features provide spatial details, but they also respond to local textures and false boundaries that can be reinforced by noisy pseudo-labels. The last-layer feature contains stronger semantic information and is less sensitive to these local errors, making it more suitable for stable subject localization.

Therefore, STI uses the last-layer feature to establish a reliable subject prior, while the student uses multi-level features to recover fine details under this constraint. In this design, the teacher does not need to produce a complete fine-grained mask. Its role is to provide a stable subject-level constraint, which allows the student to exploit spatial details more safely during the subsequent detail exploration stage.

Table 9: Effect of last-layer and multi-level features under fully supervised and unsupervised settings on COD10K-Test.

| Decoder | Input Features | $\mathcal{S}_m \uparrow$ | $\mathcal{F}_\beta^\omega \uparrow$ | $\mathcal{F}_\beta^m \uparrow$ | $\mathcal{E}_\phi^m \uparrow$ | $\mathcal{M} \downarrow$ |
|---|---|---|---|---|---|---|
| **Fully Supervised** | | | | | | |
| DBA | Last-layer feature | .841 | .781 | .807 | .923 | .025 |
| DBA | Multi-level features | .843 | .780 | .810 | .924 | .024 |
| BiRefBlock | Multi-level features | .869 | .819 | .839 | .944 | .018 |
| **Unsupervised** | | | | | | |
| Initial pseudo-labels* | | .483 | .279 | .312 | .489 | .295 |
| DBA | Last-layer feature | .784 | .675 | .692 | .869 | .036 |
| DBA | Multi-level features | .607 | .386 | .421 | .637 | .124 |
| BiRefBlock | Multi-level features | .449 | .199 | .234 | .449 | .272 |

* Initial pseudo-labels used to initialize the teacher in STI.

**Sensitivity Analysis of the ADE Temperature.** We further study the sensitivity of ADE to the temperature $T$ on COD10K-Test. As shown in Table 10, all settings with $T < 1$ outperform $T = 1$, where $T = 1$ corresponds to the no-rescaling baseline. The default setting $T = .3$ achieves the best $S_\alpha$, $F_\beta^w$, and MAE, while also obtaining competitive $F_\beta^m$ and $E_\phi^m$.

These results show that ADE consistently improves the performance over the no-rescaling baseline and does not depend on a narrowly tuned temperature value.

Table 10: Sensitivity analysis of the ADE temperature $T$ on COD10K-Test.

| **Temperature** $T$ | $\mathcal{S}_m \uparrow$ | $\mathcal{F}_\beta^\omega \uparrow$ | $\mathcal{F}_\beta^m \uparrow$ | $\mathcal{E}_\phi^m \uparrow$ | $\mathcal{M} \downarrow$ |
|---|---|---|---|---|---|
| 0.1 | .852 | .794 | **.824** | .928 | .024 |
| 0.3 **(Default)** | **.859** | **.798** | .823 | **.934** | **.023** |
| 0.5 | .852 | .790 | .819 | .921 | .025 |
| 0.7 | .847 | .787 | .818 | .916 | .025 |
| 1.0 | .843 | .782 | .819 | .907 | .026 |