SEGDIS: SEGMENT FIRST, THEN DISCRIMINATE FOR UNIFIED SALIENT AND CAMOUFLAGED OBJECT DETECTION

Anonymous authors
Paper under double-blind review

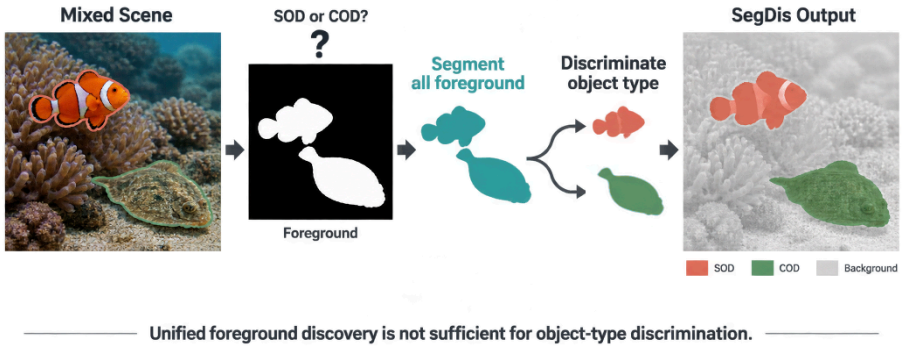


Figure 1: Foreground-conditioned pixel allocation in SEGDIS. Under the coexisting-scene setting established by USC12K (Zhou et al., 2025), a unified foreground mask alone cannot distinguish the two object types. SEGDIS estimates foreground support and allocates SOD/COD probabilities at each pixel to obtain a background/SOD/COD prediction.

ABSTRACT

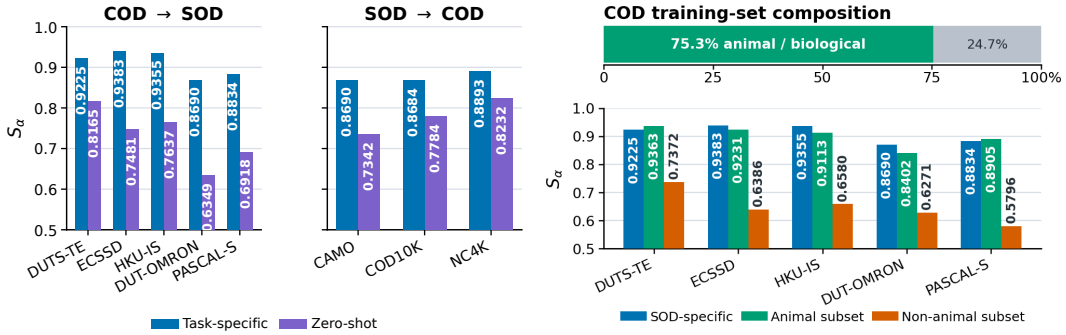
Existing joint learning approaches for salient object detection (SOD) and camouflaged object detection (COD) typically formulate both tasks as binary foreground segmentation, making it difficult to distinguish salient from camouflaged objects when they coexist. We uncover a deeper limitation through bidirectional cross-task transfer: models trained exclusively on SOD generalize surprisingly well to COD, and vice versa. This reveals that strong segmentation performance does not necessarily imply reliable object-type discrimination, suggesting that existing models primarily rely on generic foreground cues rather than task-specific characteristics of saliency and camouflage. Motivated by this observation, we propose Segment First, Then Discriminate (SegDis), a unified framework that explicitly decouples foreground discovery from object-type discrimination. SegDis first identifies task-agnostic foreground regions and then determines whether foreground pixels are salient or camouflaged by conditioning class-specific type queries on the relation between foreground and surrounding background. A subsequent uncertainty–boundary refinement stage resolves ambiguous regions and sharpens object boundaries, yielding a three-class prediction of background, salient objects, and camouflaged objects. On USC12K, SegDis achieves 71.82% FG-mIoU and 80.25% mIoU, outperforming USCNet by 3.36 and 2.22 percentage points, respectively, while reducing the Camouflage-Saliency Confusion Score (CSCS) from 7.49 to 3.27. On Scene-C, where salient and camouflaged objects coexist, SegDis obtains 73.82% IoU-S and 51.60% IoU-C, improving IoU-C by 5.87 percentage points over USCNet. These results demonstrate that decoupling foreground discovery from object-type discrimination improves unified three-class parsing and reduces salient–camouflaged confusion in mixed scenes.

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.

1 INTRODUCTION

Salient object detection (SOD) identifies objects that stand out from their surroundings, whereas camouflaged object detection (COD) recovers objects that blend into them. Although often regarded as opposing tasks, both are context-dependent and require foreground discovery. Existing research spans task-specific models, joint SOD/COD methods such as UJSC and VSCode (Li et al., 2021; Luo et al., 2024), and three-class models such as USCNet (Zhou et al., 2025). Joint binary models share foreground knowledge but cannot express different foreground types in one image. Three-class prediction addresses this output limitation, while preserving accurate foreground segmentation and reliable type allocation remains an important challenge.

Bidirectional cross-task zero-shot experiments motivate separating these capabilities. A COD-trained model achieves $S_\alpha = 0.8165$ on the SOD benchmark DUTS-TE, while an SOD-trained model reaches 0.8232 on the COD benchmark NC4K. Here $S_\alpha \in [0, 1]$ measures structural agreement with the binary foreground annotation, with higher values indicating better segmentation. Category analysis provides a further observation: the COD-trained model achieves 0.9363 on animal images in DUTS-TE and 0.7372 on non-animal images. This difference is consistent with category priors in COD training data, although it does not establish their causal effect. Together, these results show that foreground segmentation can transfer across tasks, while binary transfer scores alone do not establish reliable SOD/COD type discrimination.



(a) Bidirectional cross-task transfer.

(b) Category composition and transfer.

Figure 2: Foreground transfer across SOD and COD. The category analysis compares animal and non-animal subsets with a full-test-set SOD reference, rather than matched-subset baselines. Binary segmentation scores do not measure type discrimination.

We first establish a shared foreground pathway. Building on hierarchical FPN-style decoding, we develop a Master Decoder with residual evidence extraction and progressive refinement, forming the Progressive Foreground Refinement System (PFRS). Its foreground capability is evaluated under task-specific and joint SOD/COD training in Appendix Tables 8 and 9; the decoder control is reported separately in Appendix Table 4. However, a unified binary mask only localizes foreground and cannot assign SOD/COD types.

USC12K provides the setting needed to evaluate both capabilities, with background/SOD/COD annotations and scenes containing both foreground types (Zhou et al., 2025). In coexisting scenes, different regions require different labels, making image-level task selection insufficient. USCNet directly predicts the three classes. We instead factorize the prediction into unified foreground estimation and foreground-conditioned type allocation, allowing shared foreground support to guide pixel-wise semantic assignment.

Following this principle, we propose Segment First, Then Discriminate (SegDis). The framework first predicts a type-agnostic foreground. A pooled foreground–background semantic relation then adapts two class-specific queries, which match local features to assign conditional SOD/COD probabilities at each pixel. An uncertainty–boundary refinement stage adjusts foreground confidence, and probability composition produces the final background, SOD, and COD predictions. This design combines shared foreground modeling with explicit type allocation for unified three-class prediction.

Our contributions are summarized as follows:

- We propose SegDis, a unified framework that separates shared foreground discovery from pixel-wise SOD/COD type allocation, combining foreground support and conditional type probabilities into a three-class prediction.
- We develop relation-conditioned type queries that use foreground–background semantics to adapt SOD/COD assignment to each image, allowing coexisting foreground regions to receive different semantic labels.
- Experiments on USC12K and conventional SOD/COD benchmarks assess foreground segmentation and type discrimination. SegDis improves overall three-class segmentation and reduces saliency–camouflage confusion, with component analyses examining the contribution of its prediction design.

2 Related Work

2.1 Task-specific salient and camouflaged object detection

Task-specific SOD methods improve foreground segmentation through feature fusion, contextual modeling, and detail recovery. F3Net addresses cross-level feature discrepancies with cross-feature fusion and cascaded feedback (Wei et al., 2020); GateNet regulates encoder–decoder information flow through multi-level gating and Fold-ASPP (Zhao et al., 2020); VST models long-range dependencies with a Transformer and token-based decoding (Liu et al., 2021); and ICON integrates complementary features for salient object segmentation (Zhuge et al., 2022).

COD methods recover targets with subtle object–background differences. SINet introduces search-and-identification reasoning alongside the COD10K benchmark (Fan et al., 2020; 2022). PFNet refines ambiguous regions through positioning, focusing, and distraction mining (Mei et al., 2021), while ZoomNet learns mixed-scale representations from zoomed inputs (Pang et al., 2022). FSPNet enhances local Transformer features and adjacent-level aggregation (Huang et al., 2023); FEDER combines frequency decomposition with edge reconstruction (He et al., 2023); and FPNet and FSEL explore frequency-aware and frequency–spatial representations (Cong et al., 2023; Sun et al., 2024).

These methods provide strong references for foreground segmentation within their intended tasks. We compare our foreground pathway with them under the stated task-specific and joint training settings, while evaluating mixed-scene type allocation separately.

2.2 Joint learning of SOD and COD

Joint and general-purpose models explore shared representations for SOD and COD. UJSC connects the tasks through cross-task samples, similarity modeling, and uncertainty-aware adversarial learning (Li et al., 2021). SENet develops a general Transformer architecture with auxiliary image reconstruction (Hao et al., 2024). VSCode uses two-dimensional prompts to learn domain-specific and task-specific knowledge (Luo et al., 2024). Seamless Detection learns task-agnostic foreground representations through contrastive distillation and contextual decoding (Liu et al., 2025). Distribution-Specific Learning introduces task-specific distribution parameters within a shared network, together with saliency-based sampling (Hao et al., 2025).

These studies establish the value of sharing foreground knowledge across tasks. Their binary prediction formulations, however, do not assign distinct SOD/COD labels to coexisting foreground regions. SegDis builds on shared foreground modeling and explicitly allocates the two types at the pixel level.

2.3 Unconstrained SOD/COD detection

USCNet extends binary SOD/COD prediction to unified three-class segmentation, assigning background, salient, and camouflaged labels within a single image (Zhou et al., 2025). It combines a SAM-based architecture with attribute prompts and inter-/intra-sample queries to distinguish coexisting foreground types, moving beyond image-level task selection.

USCNet provides the closest methodological reference for our work. While it directly predicts three-class masks, SegDis explicitly separates foreground discovery from foreground-conditioned type

allocation. Foreground–background semantic relations adapt class-specific queries for pixel-wise SOD/COD assignment, and the resulting type probabilities are combined with foreground support to produce the final three-class prediction.

3 METHOD

3.1 FOREGROUND-FIRST FORMULATION

Given an image I, SEGDIS factorizes three-class segmentation into a foreground probability and a conditional SOD/COD distribution at each pixel (Figure 3). This is foreground-conditioned pixel allocation: COD is an alternative foreground label, not the complement of SOD over the whole image. Both queries learn from annotated types. The stages are trained jointly; "segment first" describes computation, not separate pretraining. No task identifier or scene label is supplied at inference.

The two decisions have different targets. Foreground supervision merges salient and camouflaged annotations. The type head performs a two-class SOD/COD decision, supervised within annotated foreground, with no background class: $\pi_{\mathrm{SOD}} + \pi_{\mathrm{COD}} = 1$. It is evaluated densely, so its maps express relative type preference across the image rather than a second foreground/background segmentation. At inference, foreground probability weights these responses, while its complement supplies the background probability in the final three-class prediction.

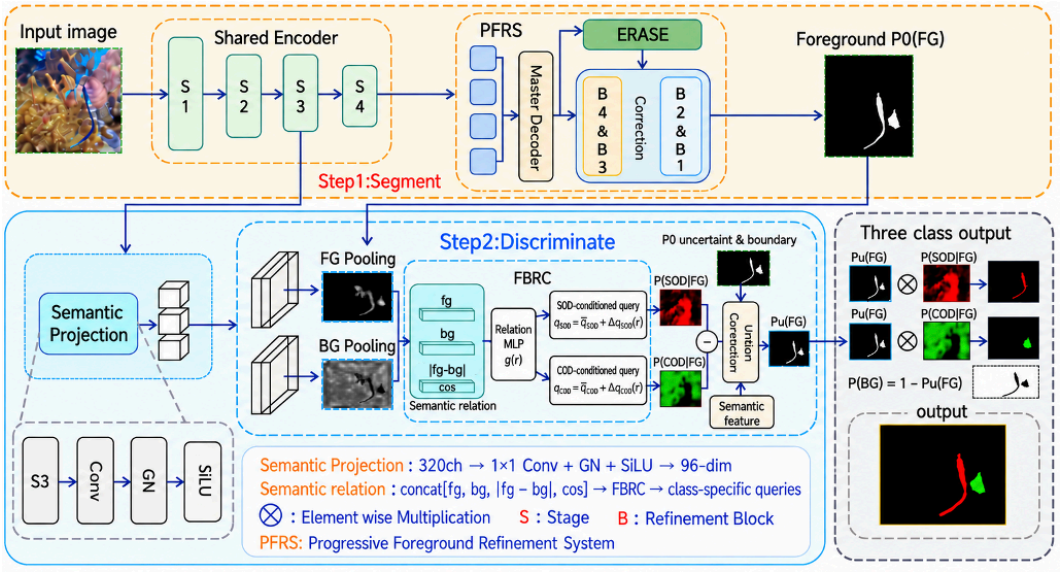


Figure 3: SEGDIS overview. PFRS estimates unified foreground. Soft pooling summarizes foreground/background semantics, and their image-level relation conditions two type queries. Local feature–query matching gives conditional SOD/COD probabilities. A bounded foreground correction precedes three-class probability composition.

3.2 PROGRESSIVE FOREGROUND REFINEMENT SYSTEM

PFRS uses a PVT-v2-B2 encoder (Wang et al., 2022) with four features $\{S_i\}_{i=1}^{4}$ of 64, 128, 320, and 512 channels. It combines a coarse prior with residual evidence and region/boundary refinement (Figure 4).

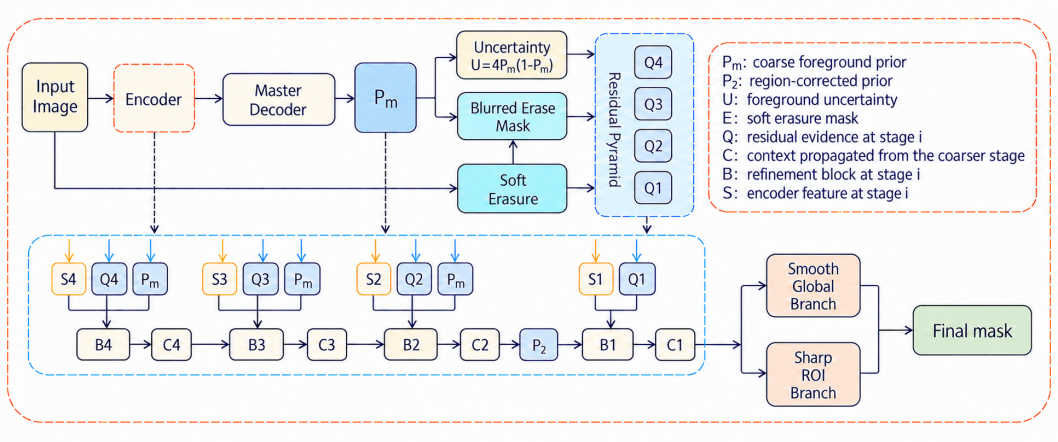


Figure 4: The PFRS pathway. Master Decoder predictions guide soft erasure and residual evidence extraction. Refinement stages B4/B3 propagate context, B2 corrects regions, and B1 recovers boundaries. These stage names denote spatial scales, not the experimental configurations in Table 2.

Coarse prior and residual evidence. The Master Decoder projects encoder features to a common dimension and fuses them top-down:

$$D_4 = \mathrm{Proj}_4(S_4), \qquad D_i = \mathrm{Fuse}_i([\mathrm{Up}(D_{i+1}), \mathrm{Proj}_i(S_i)]), \quad i = 3, 2, 1. \tag{1}$$

Here $[\cdot]$ concatenates channels, projections use 1×1 convolutions, and fusion uses 3×3 Conv–GroupNorm–SiLU. A head on D_1 produces z_M and $p_M = \sigma(z_M)$. Soft erasure replaces confident foreground with low-pass image content:

$$U_M = 4p_M(1 - p_M), \qquad E = \mathrm{Blur}(p_M^{\gamma}), \qquad I_E = I \odot (1 - E) + \mathrm{Blur}(I) \odot E. \tag{2}$$

A lightweight residual pyramid extracts multi-scale evidence $\{Q_i\}$ from $[I_E, E, U_M]$, allowing refinement to revisit content suppressed by the coarse estimate without adding a second large encoder.

Progressive correction. Coarse stages aggregate encoder features, residual evidence, foreground probabilities, uncertainty, and propagated context. At the region stage, add/remove responses update the current probability:

$$p' = \mathrm{clip}_{[0,1]}\{p + \lambda[(1 - p)\sigma(a) - p\sigma(d)]\}. \tag{3}$$

The boundary stage combines a Smooth Global Branch with a Sharp ROI Branch using ROI Align (He et al., 2017). Their corrections yield the base foreground logit z_0 and $p_0 = \sigma(z_0)$. The decoder detail and controlled Standard FPN comparison are given in Appendix Figure 11 and Table 4.

3.3 Relation-conditioned pixel-wise type allocation

Image-level foreground–background relation. A 1×1 Conv–GroupNorm–SiLU projects S3 to a 96-channel semantic map F, resized to the resolution of p_0. Foreground and background soft pooling give

$$f_{\mathrm{fg}} = \frac{\sum_x F(x)\,\mathrm{sg}[p_0(x)]}{\max(\sum_x \mathrm{sg}[p_0(x)], \epsilon)}, \qquad f_{\mathrm{bg}} = \frac{\sum_x F(x)\,\mathrm{sg}[1 - p_0(x)]}{\max(\sum_x \mathrm{sg}[1 - p_0(x)], \epsilon)}. \tag{4}$$

The stop-gradient operator sg blocks gradients through the pooling weights, while the semantic features remain trainable. The relation

$$r = [f_{\mathrm{fg}}, f_{\mathrm{bg}}, |f_{\mathrm{fg}} - f_{\mathrm{bg}}|, \cos(f_{\mathrm{fg}}, f_{\mathrm{bg}})] \in \mathbb{R}^{289} \tag{5}$$

summarizes the whole image; it is not a separate neighborhood descriptor for each object.

Independent query adaptation. Two base queries are randomly initialized and learned through SOD/COD label supervision; they are not fixed or pretrained semantic embeddings. A linear layer, LayerNorm, and SiLU map r to two independent 96-dimensional offsets:

$$[\Delta q_{\mathrm{SOD}}, \Delta q_{\mathrm{COD}}] = g(r), \qquad q_t = \bar{q}_t + \Delta q_t, \quad t \in \{\mathrm{SOD}, \mathrm{COD}\}. \tag{6}$$

For a locally projected feature $F_l(x)$, matching gives

$$\pi_t(x) = P(t \mid \mathrm{FG}, x) = \frac{\exp(q_t^\top F_l(x)/\tau)}{\sum_j \exp(q_j^\top F_l(x)/\tau)}, \qquad \tau = \max(|\tau_{\mathrm{learned}}|, 0.25). \tag{7}$$

Thus image-level context adapts a pixel-wise decision. Independent offsets are essential: an identical additive offset to both queries would cancel in the two-way softmax and leave the type probabilities unchanged.

3.4 Foreground correction and three-class supervision

The final correction uses local features, foreground uncertainty $u_0 = 4p_0(1 - p_0)$, a boundary response $b_0 = \mathrm{MaxPool}_{3\times3}(p_0) - \mathrm{MinPool}_{3\times3}(p_0)$, and the signed type preference $\pi_{\mathrm{SOD}} - \pi_{\mathrm{COD}}$:

$$\Delta z = 0.2 \tanh h([F_l, u_0, b_0, \pi_{\mathrm{SOD}} - \pi_{\mathrm{COD}}]), \qquad z_u = z_0 + \mathrm{sg}[(u_0 + b_0)/2]\Delta z. \tag{8}$$

This bounded update refines foreground confidence; it does not recompute the conditional type distribution. With $p_u = \sigma(z_u)$, the final probabilities are

$$p_{\mathrm{BG}} = 1 - p_u, \qquad p_{\mathrm{SOD}} = p_u \pi_{\mathrm{SOD}}, \qquad p_{\mathrm{COD}} = p_u \pi_{\mathrm{COD}}. \tag{9}$$

They sum to one, and a three-way argmax yields the label. Training minimizes

$$\mathcal{L} = \mathcal{L}_{\mathrm{fg}} + 0.5\mathcal{L}_{\mathrm{type}} + 0.5\mathcal{L}_{\mathrm{cal}}. \tag{10}$$

$\mathcal{L}_{\mathrm{fg}}$ is the foreground structure loss; $\mathcal{L}_{\mathrm{type}}$ is cross-entropy over ground-truth foreground pixels; $\mathcal{L}_{\mathrm{cal}}$ is negative log-likelihood of the composed three-class probabilities over all pixels. The last term provides joint output supervision, rather than a separate guarantee of confidence calibration.

4 Experiments

4.1 Protocols and evaluation

We separate three experiments. Task-specific PFRS learns binary foreground from one SOD or COD training set. Joint PFRS uses balanced alternating SOD and COD batches with a shared binary objective. Complete SegDis is trained end-to-end on USC12K with background/SOD/COD labels. The first two settings measure foreground discovery; the third measures unified three-class parsing. They use different training data and must not be interpreted as interchangeable checkpoints.

Implementation. All models use an ImageNet-pretrained PVT-v2-B2. Conventional PFRS uses 384×384 inputs; complete SegDis uses 352×352, 50 epochs, batch size 8, and AdamW with an initial learning rate of 10^{-4} and cosine decay. The complete model is initialized without a separately trained PFRS checkpoint. Appendix Table 3 details the settings.

Metrics. USC12K has 3,600 test images across SOD-only Scene-A, COD-only Scene-B, coexisting Scene-C, and background-only Scene-D (Zhou et al., 2025). Following the fixed USCNet protocol, images are bilinearly resized to 352×352, labels are resized by nearest neighbor, and three-class argmax predictions are evaluated at that resolution. We report class intersection over union (IoU), foreground mean IoU (FG-mIoU), three-class mean IoU (mIoU), mean class accuracy (mAcc), and the Camouflage-Saliency Confusion Score (CSCS). Conventional binary evaluation uses S_α, mean F_β, weighted F_β, mean E_ϕ, and mean absolute error (MAE). Foreground-only segmentation, type accuracy inside annotated foreground, and final three-class parsing are evaluated separately (Appendix Figure 7). Only the last evaluates the complete task.

4.2 FOREGROUND QUALITY AND CROSS-BENCHMARK TRANSFER

Standalone foreground pathway. Task-specific and jointly trained PFRS are evaluated separately from USC12K-trained SEGDIS. Their full conventional SOD/COD results are in Appendix Tables 8 and 9. For example, joint PFRS reaches $S_\alpha = 0.9241$ on DUTS-TE and 0.8701 on COD10K, versus 0.9225 and 0.8684 for task-specific training. Other datasets show small decreases, so shared training preserves useful foreground capability without uniformly improving it. The decoder control in Appendix Table 4 changes only the coarse decoder inside PFRS.

USC12K-trained model on conventional benchmarks. We evaluate the complete E5 checkpoint on four COD and five SOD benchmarks without fine-tuning. Appendix Table 5 measures binary quality from p_u at the original annotation resolution. Appendix Table 6 measures conditional type accuracy only inside ground-truth foreground, excluding localization errors from this diagnostic. The same checkpoint reaches $S_\alpha = 0.9058$ on DUTS-TE, 0.8183 on COD10K, and 0.8591 on NC4K. Foreground-pixel type accuracy ranges from 80.30% on PASCAL-S to 99.98% on CHAMELEON. These complementary measurements do not establish performance on entirely unseen source images without a source-overlap audit, nor do single-type benchmarks replace Scene-C evaluation.

4.3 UNIFIED PARSING ON USC12K

Table 1 presents the complete USC12K benchmark comparison. External results are reported by Zhou et al. (2025), not reproduced in our environment. The group labels identify the methods' original tasks; these external rows describe their evaluation in the USC12K benchmark, not their conventional binary test results.

Table 1: Complete comparison on the USC12K test set (%). External rows are reported by Zhou et al. (2025). SegDis uses the same test split and fixed-352 evaluation. Higher is better except for CSCS; bold highlights SegDis values that exceed all displayed external methods.

			Scene A	Scene B	Scene C		Overall scenes				
Task	**Method**	**Venue**	**IoU-S**	**IoU-C**	**IoU-S**	**IoU-C**	**IoU-S**	**IoU-C**	**mIoU**	**mAcc**	**CSCS**
SOD-specific methods											
SOD	GateNet	ECCV'20	68.32	54.26	66.85	35.03	65.08	44.17	68.27	78.07	11.30
	F3Net	AAAI'20	70.05	52.62	67.20	36.38	66.12	44.81	68.80	77.86	9.36
	MSFNet	MM'21	70.14	54.78	69.92	36.64	66.69	45.89	69.40	79.77	9.90
	VST	ICCV'21	68.14	49.82	61.61	22.56	63.18	38.45	65.55	74.77	11.30
	EDN	TIP'22	71.59	57.94	69.37	37.70	68.00	48.27	70.70	80.60	9.23
	PGNet	CVPR'22	74.69	57.31	71.94	37.21	70.72	48.78	71.82	80.76	7.71
	ICON	TPAMI'22	68.09	50.57	67.48	30.65	65.86	45.53	68.99	79.53	10.24
	MDSAM	MM'24	72.96	56.16	67.21	36.06	69.67	49.05	71.67	82.92	10.21
COD-specific methods											
COD	SINet-V2	TPAMI'21	72.96	56.16	67.21	36.06	69.50	47.47	70.20	79.58	8.83
	PFNet	CVPR'21	69.07	52.83	67.20	32.81	65.73	43.76	68.30	78.00	10.04
	ZoomNet	CVPR'22	74.11	51.12	66.79	29.69	66.43	43.28	68.35	77.72	8.88
	FEDER	CVPR'23	74.35	58.04	67.66	32.26	68.65	46.46	70.32	81.27	10.01
	PRNet	TCSVT'24	76.10	61.54	60.10	32.16	68.68	50.88	71.87	82.89	8.40
	ICEG	ICLR'24	73.67	68.38	68.43	44.33	69.22	58.71	74.68	83.53	8.16
	CamoDiffusion	AAAI'24	75.01	59.39	53.49	45.03	63.49	52.80	70.70	77.73	7.73
	CamoFormer	TPAMI'24	75.88	66.19	73.33	44.14	71.86	56.09	74.81	84.17	7.57
	PGT	CVIU'24	72.75	61.51	70.01	41.21	71.46	56.83	75.03	83.35	9.09
	SAM-Adapter	ICCVW'23	78.90	67.69	68.19	27.73	70.66	52.69	73.38	83.35	10.28
	SAM2-Adapter	arXiv'24	78.75	70.28	69.01	38.20	71.42	56.71	74.98	84.74	9.12
Unified SOD/COD methods											
Unified	EVP	CVPR'23	75.85	59.81	71.41	37.64	70.30	50.36	72.16	79.96	8.67
	VSCode	CVPR'24	76.04	60.31	72.58	39.46	71.08	55.14	74.17	84.01	8.17
	USCNet	ICCV'25	79.70	74.99	74.80	45.73	75.57	61.34	78.03	87.92	7.49
Unified	**SegDis (Ours, E5)**	**–**	**85.21**	68.44	73.82	**51.60**	**80.04**	**63.60**	**80.25**	87.46	**3.27**

SEGDIS reaches 71.82 FG-mIoU and 80.25 mIoU, exceeding USCNet by 3.36 and 2.22 percentage points, respectively, and reduces CSCS from 7.49 to 3.27. Overall IoU-S/IoU-C are 80.04/63.60. On Scene-C, IoU-C improves from 45.73 to 51.60, while IoU-S is 73.82 versus USCNet's 74.80. The gains are therefore not uniform: USCNet also retains higher Scene-B IoU-C. Our strongest evidence is improved overall parsing and reduced confusion, together with better camouflaged-region segmentation in coexisting scenes.

4.4 COMPONENT ANALYSIS

Table 2 compares designs sharing PFRS. E0 predicts two foreground channels; E1 introduces foreground factorization with learned SOD/COD queries; E2 adds foreground correction; and E3 adds composed supervision. E4 uses relation-generated queries alone, while E5 combines learned base queries with independent relation-conditioned offsets.

Table 2: Component analysis under the USCNet-352 protocol (%). Q: randomly initialized, label-trained base SOD/COD queries; Corr.: foreground correction; Calib.: composed three-class supervision; FBRC: foreground–background relation conditioning. E4 retains query matching without base vectors. Bold: column best.

	Components					USC12K metrics (%)						
Config.	PFRS	Q	Corr.	Calib.	FBRC	FG-mIoU	mIoU	IoU-S	IoU-C	CSCS↓	C-S	C-C
E0	✓					64.874	75.349	74.775	54.974	4.700	68.390	40.221
E1	✓	✓				70.867	79.591	78.255	63.480	3.508	72.086	51.755
E2	✓	✓	✓			71.088	79.748	78.722	63.454	3.399	73.144	51.449
E3	✓	✓	✓	✓		71.049	79.719	78.852	63.246	3.362	**74.017**	51.724
E4	✓		✓	✓	✓	71.629	80.101	79.284	**63.973**	3.296	71.833	**52.380**
E5 (Ours)	✓	✓	✓	✓	✓	**71.819**	**80.250**	**80.042**	63.598	**3.275**	73.819	51.605

E0→E1 changes both factorization and head, improving FG-mIoU by 5.993 points. E2 improves overall IoU; E3 reduces CSCS with a small IoU trade-off. E4 leads overall and Scene-C IoU-C; E5 leads FG-mIoU, mIoU, IoU-S, and CSCS, exceeding E3/E4 by 0.770/0.190 FG-mIoU points. Results are from single runs.

4.5 QUALITATIVE EVIDENCE

Figure 5 shows three coexisting scenes from the complete checkpoint, with ground truth alongside intermediate and final outputs. Conditional maps respond differently to regions in the same image. They may also respond outside objects because type supervision is restricted to foreground; composition suppresses these responses according to p_u. The flower scene retains local type errors, illustrating that foreground recovery does not guarantee correct type allocation. Additional visualizations are retained in Appendix Figures 8 and 9.

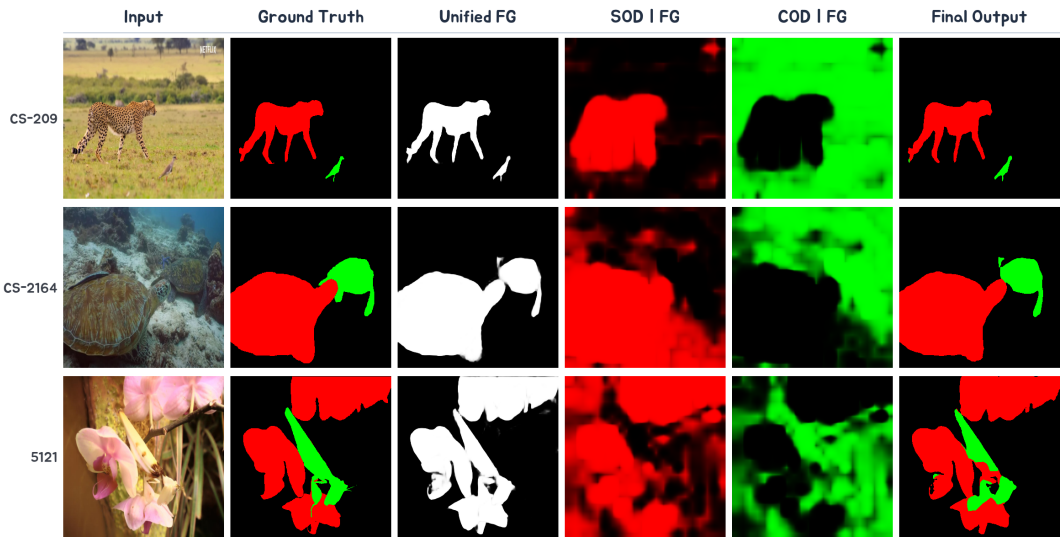


Figure 5: Coexisting-scene inference. White: foreground; red/green: SOD/COD. Unmasked type maps contain only SOD/COD probabilities; black indicates low type probability. Last column: composed background/SOD/COD prediction.

Reading conditional maps. The intermediate SOD/COD maps contain exactly two conditional probabilities, with no background channel. Their sum is one: a dark region in one map has low probability for that type and high probability for the other. They show type preference, not a second foreground/background segmentation. Foreground composition supplies the background class in the final three-class output.

Type swaps and perceptual prominence. Figure 6 shows largely preserved foreground with exchanged type labels. In some USC12K scenes, the model's SOD prediction appears visually prominent but differs from the annotated salient object. This suggests a possible mismatch between perceived saliency and dataset type annotations. Camouflaged-object size may also affect type discrimination, particularly when COD objects dominate the foreground. Appendix Figures 12–15 provide 13 further cases.

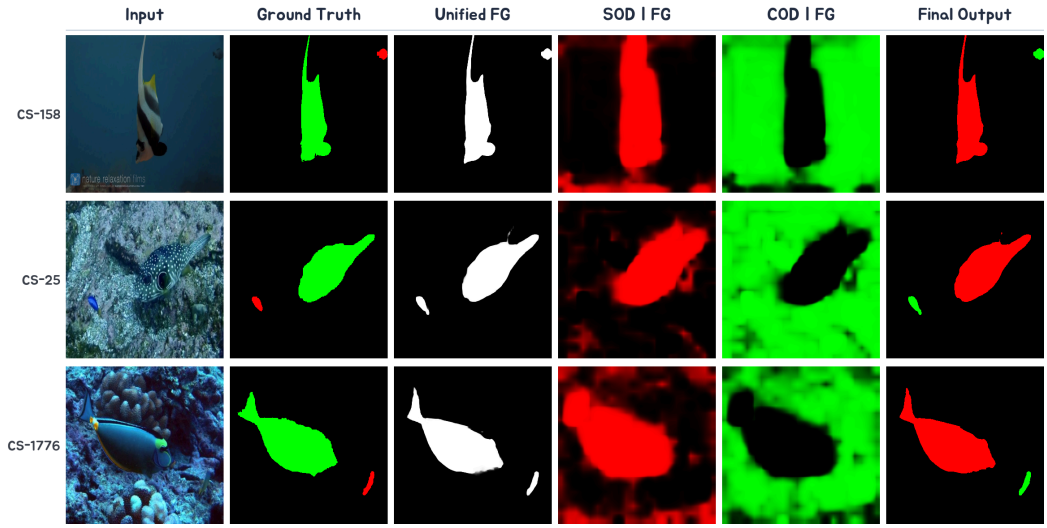


Figure 6: Selected type-swap cases (CS-158, CS-25, CS-1776). Red/green denote SOD/COD. Predicted saliency can differ from annotated saliency despite largely recovered foreground; visually prominent COD objects illustrate possible perceptual and size-related ambiguity.

Limitations. Global pooling can mix distinct objects and backgrounds. Missed foreground and incorrect types cannot always be recovered by bounded correction. Moreover, uncertain types can split foreground probability, so $p_u > 0.5$ does not guarantee a foreground label after three-class argmax. USC12K is our only mixed-scene benchmark, and seed variability remains unmeasured.

Possible annotation–perception mismatch and sensitivity to COD object size complicate interpretation of type errors. These qualitative observations require perceptual and size-controlled evaluation. Appendix Figure 10 provides additional localization and clutter failures.

5 Conclusion

SegDis uses PFRS and relation-conditioned queries to separate foreground discovery from SOD/COD assignment. Under the fixed USCNet-352 protocol, it achieves 71.82 FG-mIoU, 80.25 mIoU, and 3.27 CSCS. These results support the foreground-first factorization in the evaluated USC12K setting, while separate foreground and type diagnostics expose remaining localization and confusion errors.

AI Assistance Disclosure

Large language model tools assisted with identifying search keywords and potentially relevant studies on salient object detection, camouflaged object detection, and their joint learning. The authors checked the original papers before incorporating relevant references.

REFERENCES

Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3359–3367, 2023.

Tianrun Chen, Ankang Lu, Lanyun Zhu, Chaotao Ding, Chunan Yu, Deyi Ji, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam2-adapter: Evaluating and adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2408.04579*, 2024a.

Zhongxi Chen, Ke Sun, and Xianming Lin. Camodiffusion: Camouflaged object detection via conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1272–1280, 2024b.

Runmin Cong, Mengyao Sun, Sanyi Zhang, Xiaofei Zhou, Wei Zhang, and Yao Zhao. Frequency perception network for camouflaged object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2777–2787, 2020.

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.

Shixuan Gao, Pingping Zhang, Tianyu Yan, and Huchuan Lu. Multi-scale and detail-enhanced segment anything model for salient object detection. *arXiv preprint arXiv:2408.04326*, 2024.

Chao Hao, Zitong Yu, Xin Liu, Jun Xu, Huanjing Yue, and Jingyu Yang. A simple yet effective network based on vision transformer for camouflaged object and salient object detection. *arXiv preprint arXiv:2402.18922*, 2024.

Chao Hao, Zitong Yu, Xin Liu, Yuhao Wang, Weicheng Xie, Jingang Shi, Huanjing Yue, and Jingyu Yang. Distribution-specific learning for joint salient and camouflaged object detection. *arXiv preprint arXiv:2508.06063*, 2025.

Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22046–22055, 2023.

Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Chenyu You, Zhenhua Guo, Xiu Li, Martin Danelljan, and Fisher Yu. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. In *The Twelfth International Conference on Learning Representations*, 2024.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.

Xihang Hu, Xiaoli Zhang, Fasheng Wang, Jing Sun, and Fuming Sun. Efficient camouflaged object detection network based on global localization perception and local guidance refinement. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huan-Xiong Chen, Jie Qin, and Hailing Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10071–10081, 2021.

Jiepan Li, Wei He, Fangxiao Lu, and Hongyan Zhang. Toward complex backgrounds: A unified difference-aware decoder for binary segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):2372–2386, 2026.

Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4722–4732, 2021.

Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19434–19445, 2023.

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. *Expert Systems with Applications*, 274:126912, 2025.

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

Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8772–8781, 2021.

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

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

Rui Wang, Caijuan Shi, Changyu Duan, Weixiang Gao, Hongli Zhu, Yunchao Wei, and Meiqin Liu. Camouflaged object segmentation with prior via two-stage training. *Computer Vision and Image Understanding*, 246:104061, 2024.

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

Jun Wei, Shuhui Wang, and Qingming Huang. F3net: Fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12321–12328, 2020.

Yu-Huan Wu, Yun Liu, Le Zhang, Ming-Ming Cheng, and Bo Ren. EDN: Salient object detection via extremely-downsampled network. *IEEE Transactions on Image Processing*, 31:3125–3136, 2022.

Chenxi Xie, Changqun Xia, Mingcan Ma, Zhirui Zhao, Xiaowu Chen, and Jia Li. Pyramid grafting network for one-stage high resolution saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11717–11726, 2022.

Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. Camoformer: Masked separable attention for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Miao Zhang, Tingwei Liu, Yongri Piao, Shunyu Yao, and Huchuan Lu. Auto-MSFNet: Search multi-scale fusion network for salient object detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 667–676, 2021.

Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *European Conference on Computer Vision*, pp. 35–51, 2020.

Zhangjun Zhou, Yiping Li, Chunlin Zhong, Jianuo Huang, Jialun Pei, Hua Li, and He Tang. Rethinking detecting salient and camouflaged objects in unconstrained scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22372–22382, 2025.

Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3738–3752, 2022.

A SUPPLEMENTARY EXPERIMENTS

This section provides reproducibility details and supplementary evidence, including training protocols, decoder controls, cross-dataset transfer, foreground-level type discrimination, and failure cases. The full USC12K comparison is in main-text Table 1. Its additional baseline sources include SOD methods (Zhang et al., 2021; Wu et al., 2022; Xie et al., 2022; Gao et al., 2024), COD methods (Hu et al., 2024; He et al., 2024; Chen et al., 2024b; Yin et al., 2024; Wang et al., 2024), and prompt/adaptation methods (Liu et al., 2023; Chen et al., 2023; 2024a).

A.1 WHAT EACH EVALUATION MEASURES

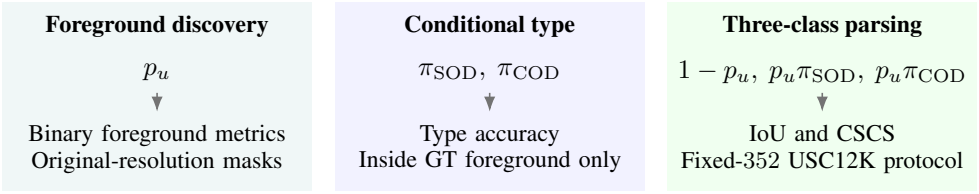


Figure 7: Evaluation outputs and protocols, not network stages. Binary foreground metrics evaluate p_u at original annotation resolution. Conditional type accuracy evaluates π only on ground-truth foreground. USC12K evaluates the composed three-class argmax at fixed 352×352. These diagnostics use different targets and are not interchangeable scores.

The conditional type head is supervised with SOD/COD labels on annotated foreground pixels. Its two probabilities sum to one even outside predicted foreground, where the conditional map is not a final semantic label. Background instead has probability $1 - p_u$. Consequently, a broad COD response in a conditional map does not mean that all non-salient background is assigned to COD in the final output. Conversely, this factorization alone does not prove that the learned representation captures a causal or context-invariant concept of camouflage.

A.2 TRAINING PROTOCOL DETAILS

Table 3: Training configurations. Task-specific and joint PFRS are binary foreground experiments; the USC12K column is the complete three-class SegDis model.

Setting	**Task-specific PFRS**	**Joint PFRS**	**SegDis on USC12K**
Training data	One SOD or COD training set	DUTS-TR + COD train	USC12K train set
Sampling	Single-task batches	Balanced SOD/COD alternation	Three-class mixed batches
Input / epochs	384×384 / 100	384×384 / 50	352×352 / 50
Prediction target	Binary foreground	Shared binary foreground	Background / SOD / COD
Objective	$\mathcal{L}_{\mathrm{fg}}$	$\mathcal{L}_{\mathrm{fg}}$	$\mathcal{L}_{\mathrm{fg}} + 0.5\mathcal{L}_{\mathrm{type}} + 0.5\mathcal{L}_{\mathrm{cal}}$
Shared settings			
Backbone	PVT-v2-B2, ImageNet initialized		
Optimizer	AdamW, cosine annealing		
Initial learning rates	Decoder 10^{-4}; backbone 10^{-5}	Decoder 10^{-4}; backbone 10^{-5}	All parameters 10^{-4}
Other settings	Batch 8; weight decay 5×10^{-4}; gradient clip 5.0; random horizontal flip; seed 42		

The complete-model optimizer uses one parameter group in the reference implementation (commit `f9aad63`); conventional PFRS uses separate backbone and decoder groups. The complete-model checkpoint is selected by epoch-average training loss. Experimental names E0–E5 in this manuscript describe configurations, not backbone sizes; the complete checkpoint retains the historical metadata

string `B4`. For conventional benchmark evaluation, the sample counts below describe the local paired image/mask collections and should not be read as a claim that every collection contains the full official benchmark split.

A.3 Additional qualitative inference views

Figure 11 shows the Master Decoder used inside the foreground path. Figures 8 and 9 provide larger views of the Scene-C examples retained from the supplied manuscript. These supplementary examples use the same displayed prediction conventions.



Figure 8: Enlarged qualitative inference cases from USC12K Scene-C (first two rows). The labels identify the input, ground truth, unified foreground, SOD response, COD response, and final three-class prediction.

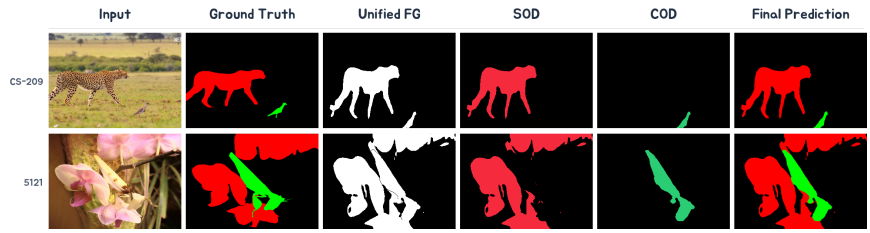


Figure 9: Enlarged qualitative inference cases from USC12K Scene-C (last two rows). The final prediction combines the foreground support with the pixel-wise SOD/COD type assignment.

A.4 Failure cases

Figure 10 reports representative errors of the current E5 checkpoint on USC12K Scene-C. The first row (CS-25) exposes a type-allocation failure relative to the annotation: the unified foreground is accurate, but the COD-labeled main object is assigned to SOD, while the SOD-labeled small object is assigned to COD. Thus, foreground discovery alone does not guarantee agreement with the annotated types. The second row (CS-2035) illustrates the other limit of the decomposition: tiny and distant foreground targets are missed by the unified foreground stage, so the later type branch has no object support to classify. The crowded CS-247 scene further shows that, even when the main SOD instance is retained, distant COD regions and clutter can still cause misses or spurious COD responses. These cases identify both remaining error sources: contextual SOD/COD confusion after foreground discovery, and unrecoverable misses when foreground discovery fails.

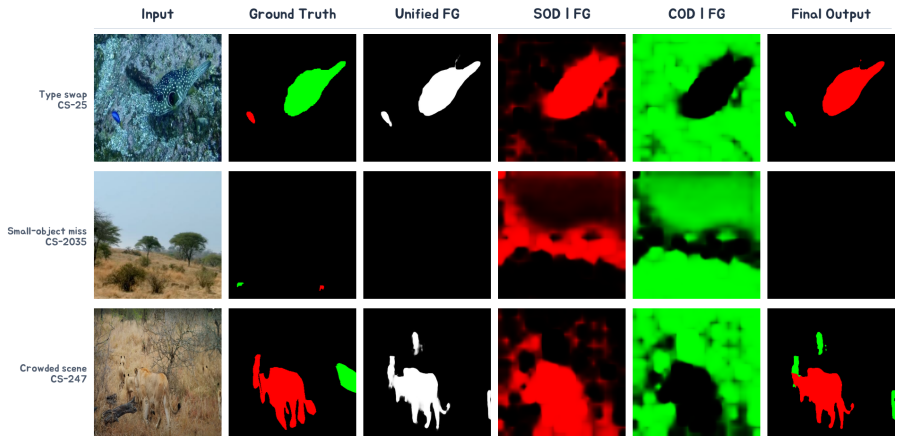


Figure 10: Additional failure analysis: type confusion (CS-25), small-target misses (CS-2035), and clutter (CS-247). Red/green denote SOD/COD.

A.5 Master Decoder and decoder control

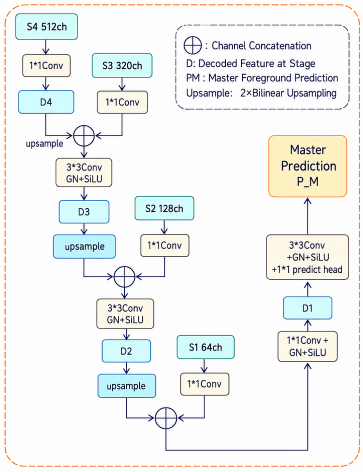


Figure 11: The Master Decoder. Encoder features are projected and fused in a top-down manner to produce the coarse foreground prediction P_M.

The decoder control changes the coarse decoder within the same PFRS pathway.

Table 4: COD comparison between Master Decoder and Standard FPN. Both use the same PVT-v2-B2 encoder and training protocol.

Dataset	Decoder	S_α ↑	$\mathbf{mF}_\beta$ ↑	$\mathbf{wF}_\beta$ ↑	$\mathbf{mE}_\phi$ ↑	MAE ↓
CAMO	Standard FPN	.8540	.8297	.7989	.9066	.0545
	Master Decoder	**.8624**	**.8436**	**.8025**	**.9188**	**.0511**
COD10K	Standard FPN	.8616	.7893	.7629	.9241	.0255
	Master Decoder	**.8644**	**.8009**	**.7769**	**.9302**	**.0236**
NC4K	Standard FPN	.8845	.8519	.8289	.9297	.0346
	Master Decoder	**.8861**	**.8600**	**.8379**	**.9341**	**.0324**

A.6 TRANSFER FROM USC12K TO CONVENTIONAL DATASETS

Table 5: Direct transfer of the USC12K-trained E5 SEGDIS to conventional COD/SOD datasets, without fine-tuning. Binary metrics are evaluated from the unified foreground probability at each image's original resolution.

Task	Dataset	#Images	S_α	meanF$_\beta$	wF$_\beta$	meanE$_\phi$	MAE↓
COD	CAMO	250	.7941	.7473	.7078	.8337	.0797
COD	CHAMELEON	76	.8591	.7956	.7761	.8847	.0364
COD	COD10K	2026	.8183	.7196	.6906	.8528	.0417
COD	NC4K	4121	.8591	.8150	.7887	.8953	.0453
SOD	DUTS-TE	5019	.9058	.8741	.8647	.9369	.0297
SOD	ECSSD	1000	.9239	.9151	.9075	.9419	.0328
SOD	HKU-IS	4447	.9263	.9115	.9063	.9497	.0245
SOD	DUT-OMRON	5168	.8448	.7719	.7591	.8673	.0507
SOD	PASCAL-S	850	.8594	.8287	.8120	.8905	.0600

A.7 TYPE DISCRIMINATION INSIDE THE FOREGROUND

Table 6: Pixel-level type accuracy of the USC12K-trained E5 SEGDIS inside the ground-truth foreground (%), without an input task identifier. This metric excludes background and is not image-level task classification.

True type	Dataset	#Images	Type accuracy
COD	CAMO	250	93.70
COD	CHAMELEON	76	99.98
COD	COD10K	2026	98.20
COD	NC4K	4121	87.19
SOD	DUTS-TE	5019	95.32
SOD	ECSSD	1000	88.84
SOD	HKU-IS	4447	84.47
SOD	DUT-OMRON	5168	97.86
SOD	PASCAL-S	850	80.30

A.8 PER-CATEGORY ANALYSIS OF THE TRANSFER SOURCE

Figure 2b presents the category analysis alongside bidirectional transfer in the Introduction. The composition estimate follows the supplied category analysis; the detailed values are retained below.

Table 7: Per-category analysis after transferring the COD-trained model to SOD. Animal and Non-animal are both COD→SOD zero-shot results; SOD-specific is the same structure trained on SOD. About 75.3% of the COD training set consists of animal or biological camouflage targets.

Dataset	Setting	S_α ↑	$\mathbf{mF}_\beta$ ↑	$\mathbf{wF}_\beta$ ↑	$\mathbf{mE}_\phi$ ↑	MAE ↓
DUTS-TE	SOD-specific/Full	.9225	.9031	.8982	.9552	.0229
	COD/Animal	.9363	.9263	.9210	.9692	.0178
	COD/Non-animal	.7372	.6887	.6336	.7646	.0897
ECSSD	SOD-specific/Full	.9383	.9421	.9366	.9617	.0246
	COD/Animal	.9231	.9214	.9122	.9428	.0293
	COD/Non-animal	.6386	.5828	.5162	.6231	.1587
HKU-IS	SOD-specific/Full	.9355	.9339	.9301	.9677	.0206
	COD/Animal	.9113	.9079	.8948	.9426	.0302
	COD/Non-animal	.6580	.6019	.5306	.6491	.1239
DUT-OMRON	SOD-specific/Full	.8690	.8186	.8083	.9014	.0425
	COD/Animal	.8402	.7868	.7612	.8755	.0385
	COD/Non-animal	.6271	.4871	.4344	.6313	.1122
PASCAL-S	SOD-specific/Full	.8834	.8704	.8581	.9242	.0497
	COD/Animal	.8905	.8855	.8753	.9306	.0428
	COD/Non-animal	.5790	.4925	.4192	.5703	.1868

B Conventional Binary Benchmark Results

The following tables provide the full conventional binary SOD/COD benchmark comparison for the PFRS foreground pathway. These experiments use the task-specific and joint protocols stated in Appendix Table 3; they complement, but do not replace, the USC12K mixed-scene evaluation in the main text.

Table 8: Comparison on conventional SOD benchmarks. Task-specific PFRS and joint PFRS use the same progressive foreground system with different training protocols. Blue shading marks PFRS variants; bold identifies the task-specific PFRS row. Dashes denote unreported metrics.

Dataset	Method	Backbone	S_α ↑	$\mathbf{mF}_\beta$ ↑	$\mathbf{wF}_\beta$ ↑	$\mathbf{mE}_\phi$ ↑	MAE↓
DUTS-TE							
	PFRS (Task-specific)	PVT-v2-B2	**.9225**	**.9031**	**.8982**	**.9552**	**.0229**
	PFRS (Joint)	PVT-v2-B2	.9241	.9025	.8977	.9556	.0228
	SENet (Hao et al., 2024)	ViT-B	.926	–	–	.953	.022
	JoNet (Hao et al., 2025)	ViT-B	.931	–	–	.958	.020
	UJSC (Li et al., 2021)	ResNet-50	.899	–	–	.937	.032
ECSSD							
	PFRS (Task-specific)	PVT-v2-B2	**.9383**	**.9421**	**.9366**	**.9617**	**.0246**
	PFRS (Joint)	PVT-v2-B2	.9290	.9294	.9218	.9501	.0299
	SENet (Hao et al., 2024)	ViT-B	.949	–	–	.968	.020
	JoNet (Hao et al., 2025)	ViT-B	.948	–	–	.966	.020
	UJSC (Li et al., 2021)	ResNet-50	.933	–	–	.967	.026
HKU-IS							
	PFRS (Task-specific)	PVT-v2-B2	**.9355**	**.9339**	**.9301**	**.9677**	**.0206**
	PFRS (Joint)	PVT-v2-B2	.9241	.9196	.9118	.9544	.0251
	SENet (Hao et al., 2024)	ViT-B	.941	–	–	.970	.019
	JoNet (Hao et al., 2025)	ViT-B	.938	–	–	.967	.019
	UJSC (Li et al., 2021)	ResNet-50	.931	–	–	.967	.026
DUT-OMRON							
	PFRS (Task-specific)	PVT-v2-B2	**.8690**	**.8186**	**.8083**	**.9014**	**.0425**
	PFRS (Joint)	PVT-v2-B2	.8628	.8050	.7934	.8945	.0437
	SENet (Hao et al., 2024)	ViT-B	.876	–	–	.899	.040
	JoNet (Hao et al., 2025)	ViT-B	.872	–	–	.895	.040
	UJSC (Li et al., 2021)	ResNet-50	.850	–	–	.884	.051
PASCAL-S							
	PFRS (Task-specific)	PVT-v2-B2	**.8834**	**.8704**	**.8581**	**.9242**	**.0497**
	PFRS (Joint)	PVT-v2-B2	.8740	.8580	.8424	.9133	.0527
	SENet (Hao et al., 2024)	ViT-B	.890	–	–	.923	.046
	JoNet (Hao et al., 2025)	ViT-B	.887	–	–	.924	.046
	UJSC (Li et al., 2021)	ResNet-50	.864	–	–	.902	.062

Table 9: Comparison on conventional COD benchmarks. Task-specific PFRS and joint PFRS use the same progressive foreground system with different training protocols. Blue shading marks PFRS variants; bold identifies the task-specific PFRS row. A dash denotes a metric not reported by the corresponding source.

Dataset	Method	Backbone	$S_\alpha \uparrow$	$mF_\beta \uparrow$	$wF_\beta \uparrow$	$mE_\phi \uparrow$	MAE↓
CAMO							
	PFRS (Task-specific)	PVT-v2-B2	**.8690**	**.8503**	**.8239**	**.9217**	**.0488**
	PFRS (Joint)	PVT-v2-B2	.8643	.8417	.8173	.9143	.0509
	FSPNet (Huang et al., 2023)	ViT-B	.8550	–	.7990	.9190	.0500
	PFNet (Mei et al., 2021)	ResNet-50	.7820	–	.6950	.8550	.0850
	SINet-V2 (Fan et al., 2022)	Res2Net-50	.8200	–	.7430	.8840	.0710
	ZoomNet (Pang et al., 2022)	ResNet-50	.8200	–	.7520	.8830	.0660
	FSEL-PVT (Sun et al., 2024)	PVT-v2-B4	.8850	–	.8510	.9420	.0400
	DAD (Li et al., 2026)	PVT-v2-B2	.8714	.8273	.7996	.9052	.0570
COD10K							
	PFRS (Task-specific)	PVT-v2-B2	**.8684**	**.8060**	**.7823**	**.9306**	**.0229**
	PFRS (Joint)	PVT-v2-B2	.8701	.8105	.7880	.9307	.0232
	FSPNet (Huang et al., 2023)	ViT-B	.8470	–	.7350	.9000	.0260
	PFNet (Mei et al., 2021)	ResNet-50	.7980	–	.6600	.8680	.0400
	SINet-V2 (Fan et al., 2022)	Res2Net-50	.8130	–	.6800	.8640	.0370
	ZoomNet (Pang et al., 2022)	ResNet-50	.8350	–	.7290	.8930	.0290
	FSEL-PVT (Sun et al., 2024)	PVT-v2-B4	.8730	–	.8000	.9280	.0210
	DAD (Li et al., 2026)	PVT-v2-B2	.8637	.7785	.7507	.9135	.0287
NC4K							
	PFRS (Task-specific)	PVT-v2-B2	**.8893**	**.8641**	**.8424**	**.9350**	**.0318**
	PFRS (Joint)	PVT-v2-B2	.8914	.8657	.8455	.9359	.0313
	FSPNet (Huang et al., 2023)	ViT-B	.8780	–	.8160	.9230	.0350
	PFNet (Mei et al., 2021)	ResNet-50	.8280	–	.7450	.8940	.0530
	SINet-V2 (Fan et al., 2022)	Res2Net-50	.8470	–	.7700	.9010	.0480
	ZoomNet (Pang et al., 2022)	ResNet-50	.8520	–	.7840	.9070	.0430
	FSEL-PVT (Sun et al., 2024)	PVT-v2-B4	.8920	–	.8530	.9410	.0300
	DAD (Li et al., 2026)	PVT-v2-B2	.8841	.8327	.8080	.9169	.0407

C Additional Type-Allocation Observations

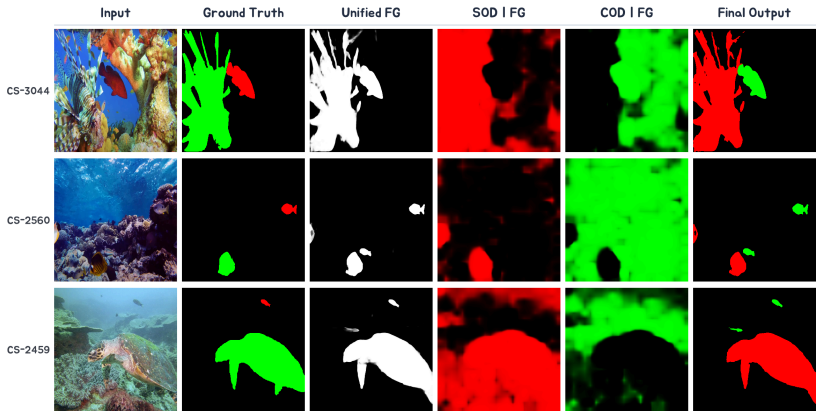


Figure 12: Additional selected cases: CS-3044, CS-2560, and CS-2459. Compare recovered foreground with the annotated and predicted types.

The following 13 cases complement the three examples in Figure 6. They are selected diagnostic examples from Scene-C, not a random sample or an estimate of benchmark error frequency. Columns show the input, dataset annotation, base foreground probability, conditional SOD/COD probabilities, and final three-class prediction. Conditional maps are unmasked. Red/green denote SOD/COD, while black in a conditional map denotes low probability rather than a third class.

Apparent perceptual prominence need not agree with the annotated type. The observed predictions may reflect relative saliency, object size, or other correlated cues; these visualizations do not identify the cause or establish annotation errors. Some examples additionally exhibit localization errors, so not all disagreements can be attributed to type allocation alone.

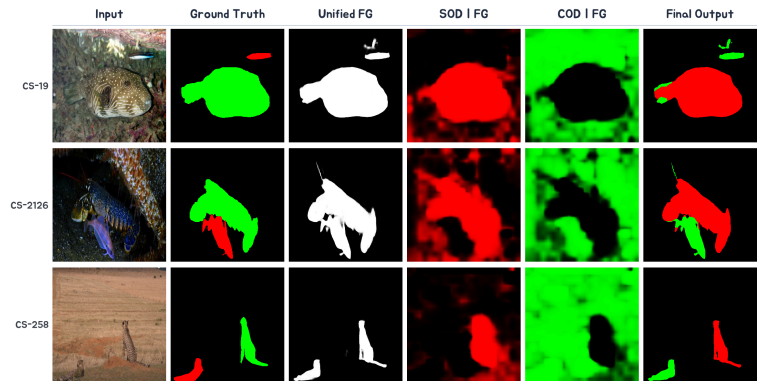


Figure 13: Additional selected cases: CS-19, CS-2126, and CS-258. The displayed type responses do not by themselves establish the cues used by the model.

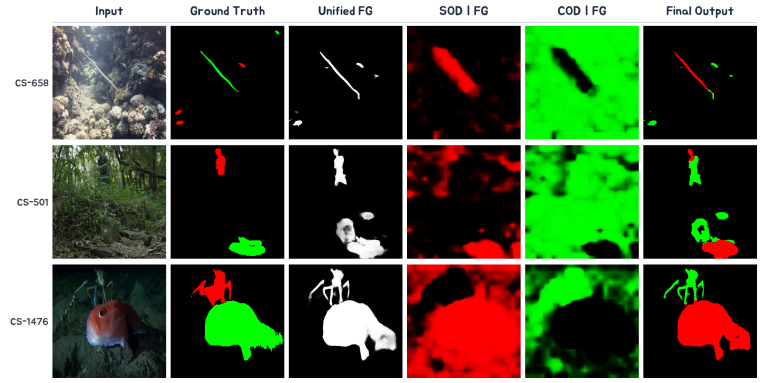


Figure 14: Additional selected cases: CS-658, CS-501, and CS-1476. Type confusion coexists with localization errors in some examples.

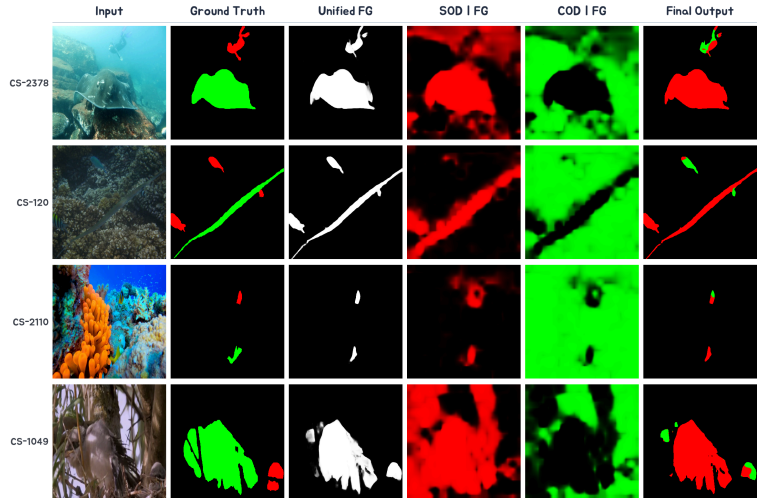


Figure 15: Additional selected cases: CS-2378, CS-120, CS-2110, and CS-1049. Disagreement with an annotation is not sufficient evidence that the model has recovered a more valid semantic label.