When Noise Meets Long-Tail: Feature-Threshold Dual Calibration for Robust Pseudo-Labeling

Ping Guo[1,2] **Zhiqi Huang**[1,3]* **Xinran Li**[1,2]

[1]Graduate School of Information, Production and Systems,
Waseda University, Kitakyushu, Japan

[2]School of Software, Dalian University of Technology, Dalian, China

[3]School of Artificial Intelligence,
The Chinese University of Hong Kong, Shenzhen, China

sr782442143@akane.waseda.jp
zhiqi.huang@akane.waseda.jp lixinran@ruri.waseda.jp

Abstract

Pseudo-labeling has become a cornerstone of learning from unlabeled data in semantic segmentation. Yet its effectiveness drops sharply in real-world scenarios where strong imaging noise and long-tailed class distributions occur together. We trace this failure to a vicious cycle of pseudo-label degradation. Imaging noise entangles foreground and background features, lowering prediction confidence across all classes, while long-tailed distributions leave tail classes with far fewer training samples and inherently lower confidence. Under fixed high-threshold filtering, these tail-class predictions are systematically filtered out, so they receive no supervision from unlabeled data and thus features keep degrading in subsequent iterations. Critically, noise and long-tail are not independent obstacles but mutually amplifying ones, and addressing either alone is insufficient. To break this cycle, we propose FTC-Seg, a Feature-Threshold dual-Calibration framework built on a standard teacher-student framework. At the feature level, Orthogonal Prototype Reconstruction (OPR) uses a set of learnable orthogonal prototypes to residually purify pixel-wise features, widening the margin between weak foreground targets and noisy backgrounds. At the threshold level, Adaptive Threshold Calibration (ATC) dynamically adjusts class-specific thresholds based on learning difficulty and prediction-distribution bias, rescuing low-confidence pseudo-labels of tail classes from systematic exclusion. Extensive experiments on four public benchmarks spanning three distinct noise modalities show that FTC-Seg achieves strong performance against state-of-the-art methods, with particularly substantial gains on tail classes. Our results establish that jointly calibrating features and thresholds is essential for robust pseudo-labeling under compounded noise and class imbalance.

The code is available at `https://github.com/pingggg516/FTC-Seg`.

1 Introduction

Pseudo-labeling, where a model generates supervision from its own confident predictions on unlabeled data, has become a cornerstone for semantic segmentation tasks under limited annotation. Methods such as FixMatch [21], UniMatch [25], and SemiVL [9] have achieved remarkable performance on standard benchmarks like Pascal VOC [4] and Cityscapes [3], all relying on a common assumption: that confident predictions filtered by a sufficiently high threshold could provide reliable supervision signals. This assumption breaks down in real-world scenarios where the data distribution departs sharply from clean and balanced benchmarks.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

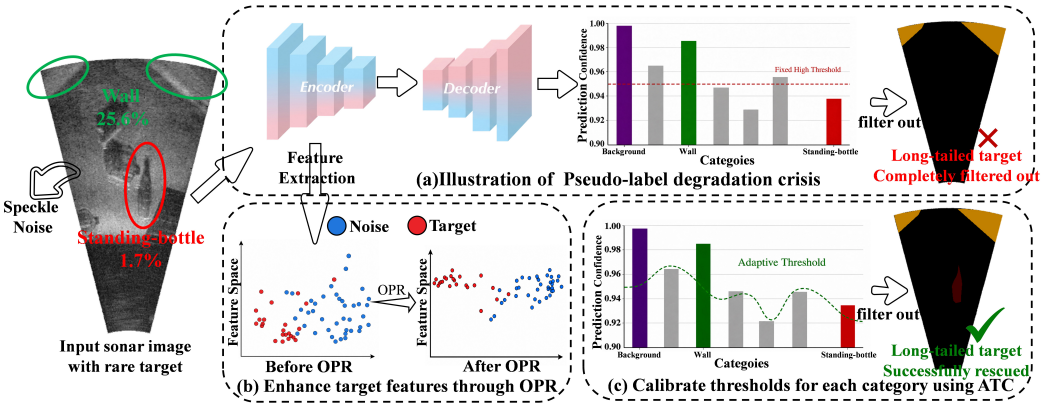


Figure 1: Illustration of the pseudo-label degradation crisis and our FTC-Seg motivation. (a) Traditional fixed high-threshold strategies are severely disturbed by speckle noise, systematically filtering out low-confidence predictions of long-tailed targets such as standing bottle. (b) At the feature level, the Orthogonal Prototype Reconstruction (OPR) enlarges the margin between target and background noise in the feature space. (c) At the threshold level, the Adaptive Threshold Calibration (ATC) dynamically lowers the threshold for rare classes, successfully rescuing long-tailed targets.

A particularly challenging regime arises when strong imaging noise and long-tailed class distributions occur together, as is common in sonar perception, underwater optical imaging, and adverse-weather driving. Environmental interference and long-tailed class distributions are also documented in the MMIO industrial defect benchmark [29]. Prior research has progressed along two largely separate directions: noise-robust pseudo-labeling [28, 19, 22, 12, 5], which targets unreliable predictions caused by image-level noise or label corruption; and long-tailed semi-supervised learning [24, 16, 23], which addresses class imbalance. However, the specific regime in which the two factors interact within the pseudo-labeling pipeline of semantic segmentation has received little explicit attention, and produces a distinct failure mode that neither line alone resolves.

We trace this failure to a vicious cycle of pseudo-label degradation, as illustrated in Figure 1. Imaging noise blurs foreground objects into the surrounding background, lowering prediction confidence across all classes, while long-tailed distributions leave tail classes with even lower confidence due to weaker representations. Under a fixed high-confidence threshold, these tail predictions are filtered out almost entirely, so tail classes receive no supervision from unlabeled data and their features keep degrading. The crucial point is that noise and long-tail are not independent obstacles but mutually amplifying ones: noise creates the confidence gap, long-tail decides who falls into it, and fixed thresholds turn the gap into permanent exclusion. Figure 2 verifies this super-linear interaction through a controlled 2×2 decoupling experiment: across FLSMD, SUIM and ACDC datasets, the tail filtered-rate under joint corruption and long-tail (CL) exceeds the linear sum of each factor alone by 1.2–2.9 percentage points. The four conditions differ in exactly one factor at a time: OB/CB share an identical balanced image list (head classes sub-sampled to match tail-class frequency) while OL/CL share an identical long-tail list, with the same pixel-level masks reused across all four cells. For ACDC, the clean Cityscapes counterpart of each image serves as the original version, providing a noise-free baseline that is unavailable for the sonar and underwater modalities.

To break this cycle, we propose FTC-Seg, which calibrates pseudo-labeling at both ends of the failure chain through Orthogonal Prototype Reconstruction (OPR) at the feature level (Figure 1(b)) and Adaptive Threshold Calibration (ATC) at the threshold level (Figure 1(c)). OPR restructures pixel features through orthogonal prototypes to make tail predictions more confident, while ATC replaces the fixed threshold with class-specific ones that relax otherwise-rejected tail predictions, together breaking the degradation cycle.

Our main contributions can be summarised as follows: **(i) A failure mechanism:** we identify and characterize a vicious cycle of pseudo-label degradation that arises from the mutually amplifying effects of imaging noise and long-tailed distributions. **(ii) A general principle:** we argue that addressing this cycle requires joint calibration at both the feature level and the threshold level, and instantiate this principle as FTC-Seg, a framework grounded in a teacher-student pseudo-labeling pipeline. **(iii) Empirical evidence across noise modalities:** through experiments on four public

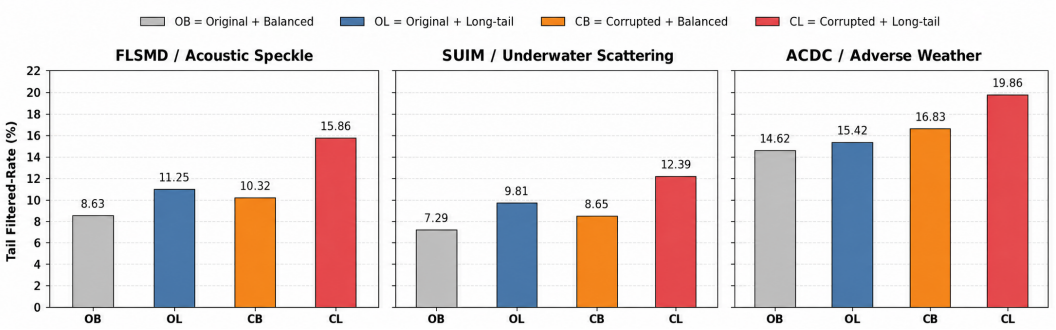


Figure 2: Controlled decoupling analysis of tail-class pseudo-label filtering. OB/OL use original images and CB/CL apply domain-aligned synthetic corruption (multiplicative speckle for FLSMD, underwater scattering for SUIM, fog for ACDC) to the same image lists, while balanced subsets are obtained by sub-sampling head classes to match tail-class frequencies; masks are unchanged across conditions. The highest filtered-rate under CL across all three modalities indicates that corruption and long-tail jointly aggravate tail-class exclusion.

benchmarks covering three distinct types of imaging noise—acoustic speckle, underwater optical scattering and adverse-weather imaging—we show that the same failure mode appears under all three noise types, and that joint feature-threshold calibration consistently resolves it.

2 Related Work

Semi-supervised semantic segmentation. Modern semi-supervised semantic segmentation builds on the teacher-student paradigm, where Mean Teacher [22] stabilizes pseudo-label generation through an exponential moving average of student weights. FixMatch [21] extends this line by introducing weak-to-strong consistency, in which pseudo-labels generated from weakly-augmented inputs are filtered by a fixed high-confidence threshold and then supervise predictions on strongly-augmented ones. Subsequent work introduces cross-network supervision [2] and and multi-level perturbations [25] for segmentation, while DTCG [13] uses complementary dual-teacher guidance for semi-supervised elliptical object detection. UniMatch V2 [26] further leverages DINOv2 as a strong backbone, while SemiVL [9] and Beyond-Pixels [8] introduce vision-language guidance and multi-scale patch-based supervision respectively. Despite their success on clean and balanced benchmarks such as Pascal VOC and Cityscapes, the fixed-threshold filtering shared by most of these methods is class-agnostic and, in noisy and long-tailed regimes, disproportionately eliminates tail-class predictions whose confidence is suppressed by both noise and limited training data.

Long-tailed semi-supervised learning. A separate line of work targets the imbalance bias of pseudo-labeling under long-tailed distributions, branching into two families. Distribution-level methods such as CReST [24] and DASO [16] for classification re-balance training by aligning pseudo-label distributions with the labeled prior or re-sampling rare classes more aggressively, but do not modify the threshold itself, so tail predictions can still be eliminated at the generation stage before any re-balancing takes effect. Threshold-level methods such as FlexMatch [27] and FreeMatch [23], though originally proposed for classification, replace the single global threshold with class-specific ones derived from the model's own learning state on unlabeled data; however, this self-referential signal cannot distinguish genuine improvement from over-prediction bias, and without an external signal to the labeled distribution, threshold adaptation can reinforce confirmation bias. More fundamentally, both families presuppose that tail-class predictions are at least separable in feature space once passed the filter—an assumption that breaks under strong imaging noise where tail features are actively confused with the background.

Noise-robust pseudo-labeling. Another line of work addresses low-quality pseudo-labels under noise through two strategies. Temporal-ensembling methods such as Mean Teacher [22] smooth short-term prediction noise via exponential moving averages, while augmentation and consistency based methods—including pseudo-label noise suppression [19] via cow-mix augmentation and noise-robust consistency regularization [28] combining multiple denoising perspectives—improve robustness through perturbation-invariant training. However, these methods operate on pseudo-labels that pass

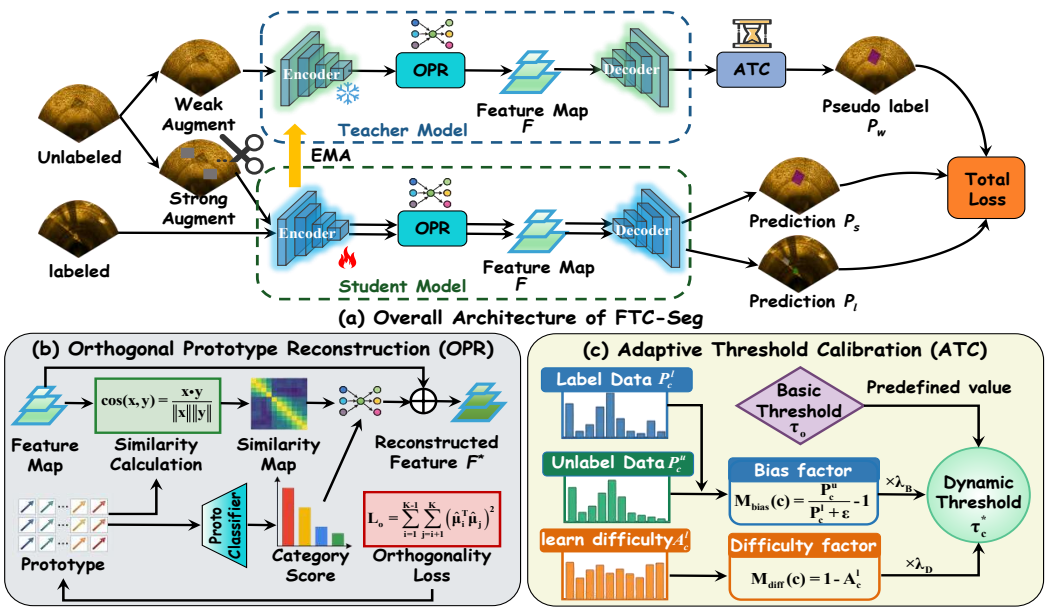


Figure 3: Overview of FTC-Seg framework. (a) Teacher-student architecture with OPR inserted between encoder and decoder, and ATC applied before pseudo-label filtering. (b) OPR computes similarities between features and learnable orthogonal prototypes, then aggregates foreground prototypes via a residual connection. (c) ATC combines a difficulty factor $\mathcal{M}_{\text{diff}}$ and a bias factor $\mathcal{M}_{\text{bias}}$ to produce class-specific thresholds.

the confidence threshold; they offer no mechanism for cases where strong imaging noise pushes tail-class predictions below the threshold and discards them before denoising can take effect, leaving tail classes systematically excluded for being low-confidence rather than for being wrong.

Our position. The three lines above each address either noise or long-tail in isolation, but the separation breaks down when the two factors occur together: threshold or distribution-level rebalancing cannot rescue predictions whose features have already been corrupted by noise, while noise-robust techniques cannot recover tail predictions filtered out before they reach denoising. Resolving this coupled failure requires intervention at both the feature level, where noise and long-tail jointly degrade representations, and the threshold level, where the resulting confidence collapse is turned into permanent exclusion.

3 Methodology

To break the degradation cycle identified in Sec 1, we propose FTC-Seg, a feature–threshold dual-calibration framework introducing two modules: Orthogonal Prototype Reconstruction (OPR) at the feature level, which purifies pixel-wise features against noise-induced entanglement, and Adaptive Threshold Calibration (ATC) at the threshold level, which replaces the class-agnostic fixed threshold with class-specific ones to prevent low-confidence tail predictions from being systematically excluded.

Overall Framework. The overall architecture of FTC-Seg is illustrated in Figure 3(a). At each training iteration, the student receives a labeled image and a strongly-augmented view of an unlabeled image, while the teacher receives the weakly-augmented view of the same unlabeled image. Both branches share an encoder–decoder, with OPR inserted between encoder and decoder to refine the feature map F. The teacher is updated as an exponential moving average (EMA) of the student and provides pseudo-labels supervising the student's predictions on the strongly-augmented view.

On the labeled data, the student is supervised by the standard pixel-wise cross-entropy loss:

$$\mathcal{L}_{sup} = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{C} Y_l^{(i,j,c)} \log P_l^{(i,j,c)}, \tag{1}$$

Table 1: Quantitative comparison of mIoU (%) on the FLSMD and FSSG datasets. The best results are shown in **bold**, and the second best results are shown in *italic*. Results of FTC-Seg and all comparative methods are averaged over 3 random seeds (±std).

Method	Venue	Encoder	FLSMD			FSSG		
			2%	5%	10%	2%	5%	10%
Labeled Only	–	RN-101	46.47 ±1.17	55.15 ±0.47	61.94 ±0.83	17.64 ±1.35	32.23 ±0.66	52.27 ±1.01
Labeled Only	–	DINOv2-S	51.08 ±0.50	61.02 ±0.95	68.48 ±0.70	35.61 ±0.89	41.67 ±1.42	54.58 ±0.32
AEL [10]	NIPS'21	RN-101	52.70 ±1.54	64.89 ±1.24	*70.73* ±0.88	53.51 ±1.02	57.84 ±0.98	63.08 ±0.46
Dual Teacher [15]	NIPS'23	MiT-B2	*56.23* ±0.70	*66.09* ±0.56	69.61 ±0.71	58.31 ±0.80	60.16 ±0.93	*65.91* ±1.01
UniMatch [25]	CVPR'23	RN-101	50.70 ±1.47	56.14 ±0.68	61.66 ±1.34	58.39 ±0.88	60.87 ±0.70	63.15 ±1.27
Beyond-Pixels [8]	ECCV'24	RN-101	48.19 ±0.60	55.52 ±1.32	60.49 ±0.93	57.20 ±1.08	*62.37* ±0.88	64.52 ±0.36
SemiVL [9]	ECCV'24	RN-101	53.38 ±1.16	65.16 ±0.89	70.02 ±0.49	*58.85* ±0.70	62.19 ±0.29	65.24 ±0.97
VC [1]	TPAMI'24	RN-101	50.53 ±1.63	59.36 ±1.31	66.86 ±1.34	49.76 ±1.20	54.75 ±0.88	59.27 ±1.09
FTC-Seg (Ours)	–	RN-101	**58.33 ±0.94**	**67.20 ±0.77**	**71.01 ±0.45**	**60.49 ±0.82**	**64.35 ±0.61**	**70.13 ±1.14**
UniMatch V2 [26]	TPAMI'25	DINOv2-S	57.24 ±0.49	66.49 ±0.51	69.81 ±0.57	58.78 ±1.32	61.21 ±0.72	64.41 ±0.75
FTC-Seg (Ours)	–	DINOv2-S	**62.55 ±1.22**	**68.91 ±0.69**	**72.54 ±0.72**	**63.20 ±0.98**	**65.49 ±1.07**	**67.53 ±0.68**

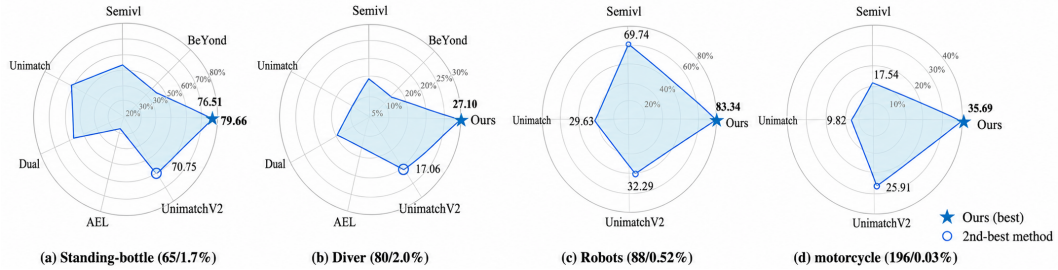


Figure 4: Performance comparison on the rarest or second-rarest long-tailed categories from the FLSMD, FSSG, SUIM and ACDC datasets under the 2% labeled-data setting.

where H and W are the spatial dimensions, C denotes the number of classes, $Y_l^{(i,j,c)}$ is the one-hot ground-truth and $P_l^{(i,j,c)}$ is the student's predicted probability for class c at pixel (i, j).

On the unlabeled data, the teacher's prediction $\tilde{P}_w$ on the weakly-augmented view is converted into a one-hot pseudo-label P_w through class-specific confidence filtering:

$$P_w^{(i,j,c)} = \begin{cases} 1, & c = \arg\max_{c'} \tilde{P}_w^{(i,j,c')} \wedge M^{(i,j)} = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

$$M^{(i,j)} = \begin{cases} 1, & \max_c \tilde{P}_w^{(i,j,c)} > \tau_c^*, \\ 0, & \text{otherwise}. \end{cases} \tag{3}$$

The class-specific threshold τ_c^* is the central quantity threshold-level calibration acts upon; its construction is detailed in Sec 3.3. The pseudo-label P_w then supervises the student's prediction P_s on the strongly-augmented view through an unsupervised consistency loss:

$$\mathcal{L}_{unsup} = -\frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{c=1}^{C} P_w^{(i,j,c)} \log P_s^{(i,j,c)}. \tag{4}$$

The student minimizes $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{unsup}}$, while the teacher updates via EMA without backpropagation. OPR refines features before decoding; ATC sets τ_c^* before pseudo-label filtering.

Orthogonal Prototype Reconstruction (OPR). At the feature level, the degradation cycle manifests as noise-induced entanglement: under strong imaging noise, pixel-wise features of foreground targets are pulled toward the surrounding background, suppressing their prediction confidence, a feature-level intervention must therefore separate foreground from background before the decoder. We argue that the intervention must satisfy three properties simultaneously: **(i)** anchoring each local pixel feature to a global signal stable against local noise; **(ii)** selectively admitting only the foreground portion, since mixing in background would re-introduce the entanglement; **(iii)** preserving the original feature rather than replacing it, since semantic segmentation depends on fine-grained spatial detail.

Table 2: Quantitative comparison of mIoU (%) on the SUIM and ACDC datasets.The best results are shown in **bold**, and the second best results are shown in *italic*. Results of FTC-Seg and all comparative methods are averaged over 3 random seeds (±std).

Method	Encoder	SUIM			ACDC		
		2%	5%	10%	2%	5%	10%
Labeled Only	RN-101	35.64 ±1.44	45.47 ±1.52	52.68 ±0.97	36.62 ±1.13	44.43 ±0.91	48.47 ±1.27
Labeled Only	DINOv2-S	56.61 ±0.80	62.56 ±1.29	64.26 ±1.15	52.60 ±0.45	56.57 ±0.62	66.98 ±0.97
UniMatch	RN-101	47.86 ±2.17	49.17 ±1.98	55.01 ±1.71	45.09 ±2.36	50.81 ±1.49	51.66 ±1.20
SemiVL	RN-101	55.99 ±0.79	58.43 ±0.60	60.15 ±1.32	45.13 ±0.57	50.65 ±0.71	52.97 ±0.97
FTC-Seg (Ours)	RN-101	54.71 ±0.53	59.05 ±1.46	62.85 ±0.97	45.56 ±1.35	51.80 ±0.89	55.73 ±1.04
UniMatch V2	DINOv2-S	*58.37* ±0.92	*66.12* ±1.07	*68.43* ±0.85	*59.59* ±2.48	*67.50* ±1.63	*68.55* ±1.77
FTC-Seg (Ours)	DINOv2-S	**66.91** ±1.35	**69.86** ±0.98	**70.21** ±0.90	**61.68** ±1.42	**69.82** ±0.97	**70.09** ±0.86

Table 3: Ablation studies of the proposed OPR and ATC components.

(a) Ablation study of OPR components.

Method	Component			Pseudo-Label Acc. (%)		Global
	Res.	$\mathcal{L}_o$	r_k	Global	Tail	mIoU
Baseline	–	–	–	67.03	56.56	58.75
w/o Residual	✗	✓	✓	68.69	**89.61**	56.93
w/o $\mathcal{L}_o$	✓	✗	✓	71.30	81.04	59.26
w/o r_k	✓	✓	✗	64.32	83.16	59.84
OPR	✓	✓	✓	**72.33**	87.80	**62.55**

(b) Quantitative analysis of ATC components.

Method	Tail Classes			Head Classes	Global
	Recall	Precision	IoU	IoU	IoU
Baseline	62.50	77.62	49.97	72.77	58.75
+ M_{diff}	77.11	79.20	62.61	73.54	60.42
+ M_{bias}	76.29	79.67	62.04	73.01	59.79
ATC	**86.40**	**84.00**	**73.87**	**75.37**	**61.90**

OPR realizes these three properties through a small set of learnable vectors called *prototypes*, shared across the dataset and converging to recurring semantic patterns. These global prototypes provide the stable signal required by property (i).

As shown in Figure 3(b), we maintain K learnable prototypes $P = \{\mu_1, \mu_2, \ldots, \mu_K\} \in \mathbb{R}^{K\times C}$, where C is the channel dimension of the encoder feature map $F \in \mathbb{R}^{C\times H\times W}$. Property (i) is realized by measuring the cosine similarity between each pixel feature $F_{(h,w)}$ and each prototype μ_k:

$$S^k_{h,w} = \frac{F^\top_{(h,w)}\mu_k}{\left\|F_{(h,w)}\right\|_2 \left\|\mu_k\right\|_2}, \tag{5}$$

where $S^k_{h,w}$ expresses how each locally noisy pixel aligns with each globally learned semantic pattern. Property (ii) requires identifying which prototypes have converged to foreground semantics, obtained from the segmentation head whose weights encode class-level decision boundaries:

$$Z^k_c = W^\top_c \mu_k + b_c, \tag{6}$$

where W_c and b_c are the segmentation head's weight and bias for class c. A suppression factor γ_k then admits only prototypes whose dominant class lies in the foreground class set $\mathcal{C}_{fg}$:

$$\gamma_k = \begin{cases} 1, & \arg\max_c Z^k_c \in C_{fg}, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Without this gating, background-aligned prototypes would contaminate the reconstruction, reinstating the entanglement OPR aims to suppress. Property (iii) is realized through a residual aggregation:

$$F^*_{(h,w)} = F_{(h,w)} + \sum_{k=1}^{K} \left(\gamma_k \cdot S^k_{h,w}\right) \mu_k. \tag{8}$$

The original feature $F_{(h,w)}$ is kept in full, while the prototype aggregation acts as a corrective term strengthening foreground-aligned components without overwriting spatial detail.

However, the three properties presuppose that the prototypes are distinct. If multiple prototypes collapse to similar directions, they produce nearly identical responses in Eq. 5 and contribute redundantly to Eq. 8, causing the prototype set to degenerate into a few repeated patterns rather than a diverse collection. We therefore impose an orthogonality regularizer on the prototype matrix:

$$\mathcal{L}_O = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \left(\hat{\mu}_i^\top \hat{\mu}_j\right)^2, \tag{9}$$

Table 4: Component-wise pseudo-label analysis on global, head, and tail classes.

Method	Pseudo-Label Acc (%)			Filtered-Rate (%)			Correctly Kept GT Rate (%)		
	Global	Head	Tail	Global	Head	Tail	Global	Head	Tail
Baseline	67.03	73.06	56.56	13.15	9.50	17.83	58.22	67.04	46.47
+ OPR	72.33	74.08	87.80	9.04	7.04	11.24	63.98	71.34	77.93
+ ATC	69.73	79.60	88.65	11.63	10.38	10.39	64.80	68.13	79.44
FTC-Seg (ours)	**73.83**	**84.13**	**89.03**	**8.26**	**6.75**	**8.43**	**67.15**	**78.21**	**80.61**

Table 5: Ablation studies of OPR layers and FTC-Seg components.

(a) Effect of applying OPR to multiple encoder stages.

OPR Layers	mIoU	Params (M)	FLOPs (G)	Infer. Time (ms)	Peak Mem. (MB)
0 (Baseline)	58.75	24.79	56.88	12.97	14279
1	61.92	24.80	56.90	13.15	14348
2	**62.55**	24.81	56.92	13.27	14398
3	61.47	24.82	56.93	13.43	14449
4	60.98	24.84	56.95	13.58	14497

(b) Components ablation analysis of FTC-Seg.

Model	EMA	OPR	ATC	mIoU
M1	–	–	–	35.61
M2	✓	–	–	58.75
M3	✓	✓	–	62.55
M4	✓	–	✓	61.90
M5	✓	✓	✓	**63.20**

penalizing squared cosine similarities between distinct prototype pairs and forcing them to spread out as a discriminative set.

Table 3(a) verifies this design: removing the residual connection collapses global mIoU from 62.55% to 56.93% though tail-class pseudo-label accuracy stays high, confirming that residual aggregation protects the spatial detail required for segmentation; removing the orthogonality loss reduces global and tail pseudo-label accuracy, confirming that prototype diversity makes the global signal discriminative; and removing the suppression factor causes the largest pseudo-label accuracy drop, confirming that without semantic gating background prototypes contaminate the reconstruction.

Adaptive Threshold Calibration (ATC). At the threshold level, the degradation cycle manifests as exclusion: once feature-level entanglement suppresses tail-class confidence, a class-agnostic fixed threshold rejects these predictions wholesale, denying tail classes any supervision and locking the cycle. A threshold-level intervention must replace the global threshold with class-specific ones that respond to each class's learning state. Self-referential adaptive thresholds estimate this state solely from the model's confidence on unlabeled data, but such a signal cannot distinguish learning progress from over-prediction bias. We argue that the intervention must combine two signals: **(i)** an internal signal reflecting how well the model has learned each class, measured against ground truth on labeled data; **(ii)** an external signal comparing the model's predicted class distribution on unlabeled data against the class prior on labeled data, exposing over or under-prediction bias that the internal signal cannot detect. ATC realizes this dual-signal design.

As shown in Figure 3(c), we maintain three global statistics for each class c: the ground-truth class distribution P_c^l and per-class learning accuracy A_c^l on labeled data, the predicted class distribution P_c^u on unlabeled data. To stabilize these statistics against mini-batch fluctuations—particularly severe for tail classes with small per-batch counts—we update them with EMA:

$$P_c^l \leftarrow \eta_m P_c^l + (1-\eta_m)\tilde{P}_c^l, \tag{10}$$

$$P_c^u \leftarrow \eta_m P_c^u + (1-\eta_m)\tilde{P}_c^u, \tag{11}$$

$$A_c^l \leftarrow \eta_m A_c^l + (1-\eta_m)\tilde{A}_c^l, \tag{12}$$

where $\tilde{P}_c^l$, $\tilde{P}_c^u$ and $\tilde{A}_c^l$ are the current-batch statistics, and η_m is the EMA momentum (set to 0.999).

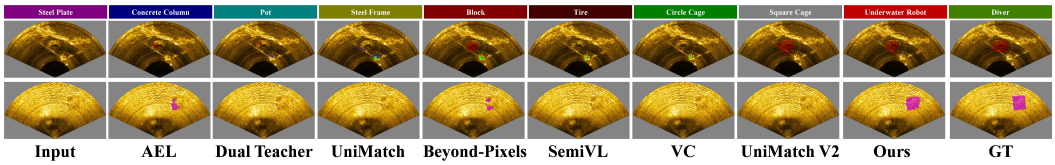
The internal signal is realized by the difficulty factor, defined as the gap between perfect accuracy and the model's actual accuracy on labeled data:

$$\mathcal{M}_{\text{diff}}(c) = 1 - A_c^l, \tag{13}$$

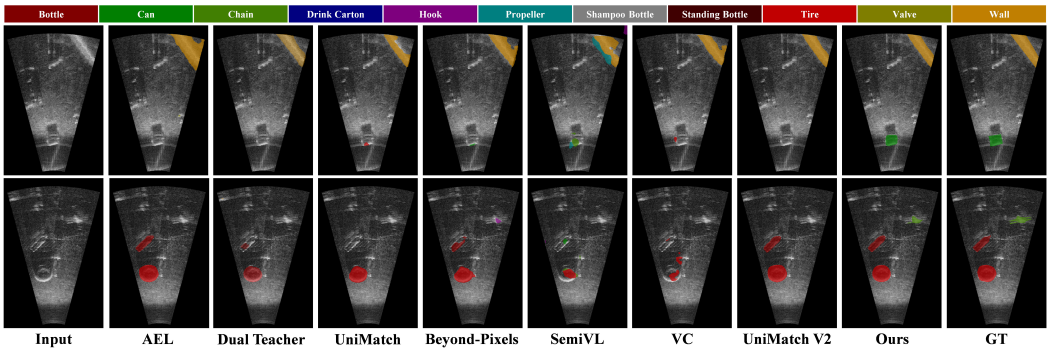
so that a poorly-performing class is identified as difficult and granted a lower threshold to admit more of its low-confidence pseudo-labels. The external signal is realized by the bias factor, contrasting the

Table 6: Dominant-class responses of 16 OPR prototypes on FSSG with 2% labels. BL/CG/SF/CC/SP/PO/SQ/TI/UR/DV denote Block/Circle Cage/Steel Frame/Concrete Column/Steel Plate/Pot/SquareCage/Tire/Underwater Robot/Diver.

Prototypes 1–4	**Prototypes 5–8**	**Prototypes 9–12**	**Prototypes 13–16**
μ_1: SP→SP→TI P=14.8, F=63.6, H=99.9	μ_5: CC→CC→DV P=16.1, F=39.8, H=99.6	μ_9: CC→UR→CC P=13.6, F=56.4, H=99.8	μ_{13}: SP→SP→CG P=15.3, F=53.2, H=99.8
μ_2: CC→DV→BL P=13.2, F=53.6, H=99.8	μ_6: SF→SF→UR P=18.0, F=79.2, H=99.7	μ_{10}: SF→SF→PO P=16.5, F=33.9, H=99.4	μ_{14}: PO→PO→CC P=17.1, F=60.8, H=99.7
μ_3: SQ→UR→DV P=14.9, F=29.7, H=99.2	μ_7: PO→PO→DV P=15.4, F=43.0, H=99.6	μ_{11}: DV→DV→PO P=14.0, F=34.2, H=99.8	μ_{15}: SF→TI→CC P=12.7, F=51.6, H=99.9
μ_4: CC→BL→CC P=17.6, F=66.4, H=99.6	μ_8: PO→PO→SP P=16.4, F=76.4, H=99.7	μ_{12}: SQ→SQ→PO P=14.4, F=37.7, H=99.4	μ_{16}: TI→TI→PO P=13.5, F=54.3, H=99.9



(a) Qualitative segmentation results on the FSSG dataset with 2% labeled data.



(b) Qualitative segmentation results on the FLSMD dataset with 2% labeled data.

Figure 5: Qualitative segmentation results compared with state-of-the-art methods on the FSSG and FLSMD datasets with 2% labeled data.

predicted class distribution on unlabeled data against the true class prior on labeled data:

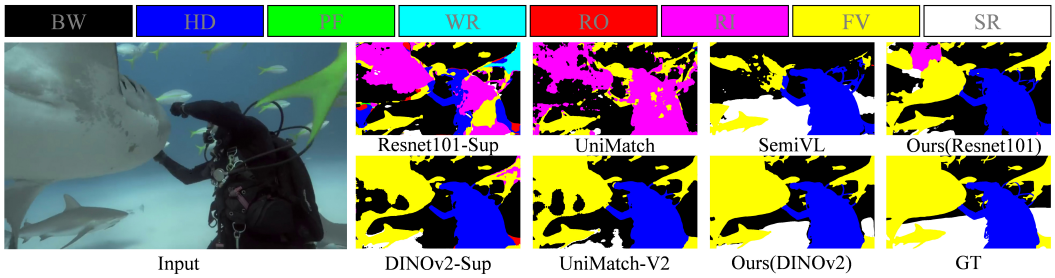
$$\mathcal{M}_{\text{bias}}(c) = \frac{P_c^u}{P_c^l + \epsilon} - 1, \tag{14}$$

where ϵ is a small constant for numerical stability. A positive $\mathcal{M}_{\text{bias}}(c)$ indicates over-prediction; a negative value indicates under-prediction. This is the signal that self-referential thresholds lack: without the labeled prior P_c^l as an anchor, a class can be heavily over-predicted yet still appear confident to the model itself. The two signals are linearly combined with the basic threshold τ_0 to form the class-specific dynamic threshold:

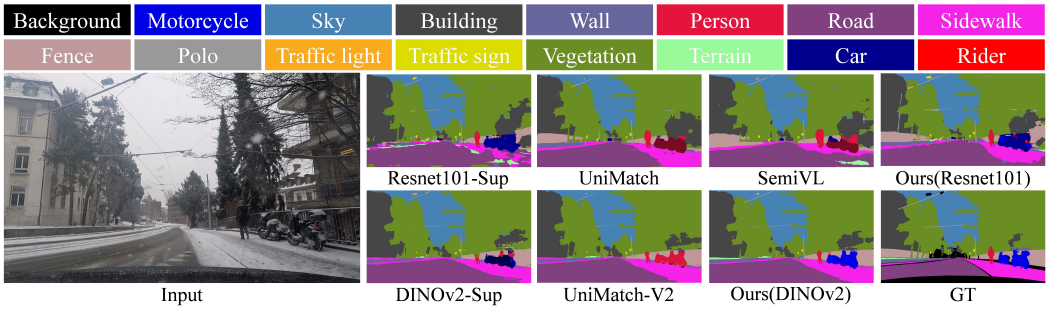
$$\tau_c^* = \tau_0 - \lambda_D \cdot \mathcal{M}_{\text{diff}}(c) + \lambda_B \cdot \mathcal{M}_{\text{bias}}(c). \tag{15}$$

where λ_D and λ_B control the influence of each signal. The signs are deliberate: difficult classes (large $\mathcal{M}_{\text{diff}}$) lower the threshold to admit more pseudo-labels for under-learned classes, while over-predicted classes (positive $\mathcal{M}_{\text{bias}}$) raise the threshold to suppress false positives that would otherwise compound confirmation bias.

Table 3(b) verifies that the two signals identify different sets of classes to rescue. Adding $\mathcal{M}_{\text{diff}}$ alone raises tail Recall to 77.11% by lowering thresholds for poorly-learned classes; adding $\mathcal{M}_{\text{bias}}$ alone raises Recall to 76.29%, but by lowering thresholds for classes under-predicted relative to the labeled prior—two criteria that overlap but do not coincide. When both signals are active, Recall jumps to 86.40%, far above either individual gain, because classes flagged by only one signal are now rescued as well. Precision rises from 77.62% to 84.00%, and tail IoU climbs from 49.97% to 73.87%, showing that the two signals expose different facets of threshold-level failure—learning-state weakness and distribution-prior bias—rather than redundantly strengthening the same effect.



(a) Qualitative segmentation results on the SUIM dataset with 2% labeled data.



(b) Qualitative segmentation results on the ACDC dataset with 2% labeled data.

Figure 6: Qualitative segmentation results compared with state-of-the-art methods on the SUIM and ACDC datasets with 2% labeled data.

4 Experiments

Experimental Setup. We evaluate FTC-Seg on four public benchmarks spanning three imaging-noise modalities: FLSMD [20] and FSSG [6] for acoustic speckle, SUIM [11] for underwater optical scattering, and ACDC [18] for adverse-weather urban driving. All exhibit long-tailed class distributions: Standing-bottle in FLSMD, Diver in FSSG, Robots (RO) in SUIM, and Rider in ACDC account for only 1.67%, 2.04%, 0.52%, and 0.02% of pixels, respectively, supporting evaluation under combined noise and class imbalance. We sample 2%, 5%, and 10% labeled data, ensuring at least one labeled instance per class, and treat the remainder as unlabeled. We use ResNet-101 [7] and DINOv2-S [17] encoders for fair comparison with prior work, reporting mean Intersection over Union (mIoU) over three random seeds. All models train for 240 epochs with AdamW [14] (lr 5×10^{-6}, weight decay 0.01, polynomial decay power 0.9) on one NVIDIA RTX 3090. We compare against AEL [10], Dual Teacher [15], UniMatch [25], Beyond-Pixels [8], SemiVL [9], VC [1], and UniMatch V2 [26], reproduced from public code under identical settings.

Main Results. Tables 1 and 2 report the comparison on FLSMD/FSSG and SUIM/ACDC respectively. FTC-Seg achieves the best or highly competitive mIoU in most configurations, with the most consistent advantages under low-label and tail-class regimes. Under the most challenging 2% setting, FTC-Seg with DINOv2-S exceeds UniMatch V2 by 5.31% mIoU on FLSMD, 4.42% on FSSG, 8.54% on SUIM, and 2.09% on ACDC. This lead being preserved across three completely different noise modalities indicates that FTC-Seg corrects a failure mode generic to noisy and long-tailed pseudo-labeling, rather than an artifact tuned to any single domain.

The advantage is pronounced on long-tailed classes. Figure 4 reports IoU on the rarest or second-rarest class of each dataset: Standing-bottle in FLSMD, Diver in FSSG, Robots in SUIM, and motorcycle in ACDC—we use the second-rarest for ACDC because more than half of the methods produce near-zero IoU on Rider. FTC-Seg attains the highest IoU on every class, with leads of 8.91%, 10.04%, 13.60%, and 9.78% over the second-best method, confirming that the framework rescues the systematic exclusion produced by the degradation cycle wherever noise and long-tail co-occur.

Figures 5 and 6 present qualitative results across all four datasets, corroborating the quantitative findings. FTC-Seg recovers the complete object extent with sharper boundaries and fewer false positives, while comparison methods fragment or miss targets hidden in noisy backgrounds.

Pseudo-Label Quality on Head and Tail Classes. Table 4 provides evidence that FTC-Seg breaks the degradation cycle of Section 1. Two observations stand out. First, the baseline's filtering is unequal: 17.83% of tail predictions are filtered out, nearly twice the 9.50% head rate—a manifestation of the claim that fixed thresholds disproportionately exclude tail classes. Under FTC-Seg, this gap collapses to 8.43% versus 6.75%, showing the disproportion is not merely reduced but largely equalized. Second, the asymmetry appears in pseudo-label accuracy: the 16.50% head–tail gap is not only closed but reversed, with tail reaching 89.03% against 84.13% on head.

The two modules intervene at different levels: OPR purifies target features so tail predictions become more confident, dropping the tail filtered-rate from 17.83% to 11.24%; ATC directly lowers the threshold for tail classes, reaching a similar 10.39%. Each mechanism substantially relieves the systematic exclusion, and combining them further reduces the rate to 8.43%.

The same pattern holds at the framework level. Decomposing FTC-Seg in Table 5(b), adding the EMA teacher–student framework to the labeled-only baseline (M1 $\rightarrow$ M2) already raises mIoU from 35.61% to 58.75%, confirming pseudo-labeling itself is the dominant source of gain when labels are scarce. On this foundation, OPR (M3) and ATC (M4) yield 62.55% and 61.90% respectively, and combining both (M5) reaches 63.20%.

Prototype semantics. Table 6 examines 16 OPR prototypes on FSSG with 2% labels. Arrows track each prototype's dominant foreground class from early to middle to final training. Final-stage statistics report class purity (P), foreground selectivity (F), and normalized foreground spatial entropy (H), all in percent. Changing dominant classes and low purity (12.7–18.0%) indicate that prototypes are not tied to fixed class identities. High spatial entropy (99.2–99.9%) and varying foreground selectivity (29.7–79.2%) suggest spatially distributed responses with different foreground preferences, consistent with class-agnostic feature reconstruction.

Computational Efficiency. Table 5(a) varies how many of the encoder's deepest stages are equipped with OPR: 0 means no OPR, 1 applies OPR only to the deepest stage, and 2 corresponds to our default architecture, where OPR is applied to the two deepest encoder stages. Higher values progressively extend OPR to shallower stages. Across all settings, parameters, FLOPs, inference time, and peak memory stay essentially flat—deviations from the no-OPR baseline are below 0.05M, 0.07G, 0.61ms, and 218MB respectively—so OPR is practically free to deploy regardless of placement. Performance peaks at two stages and degrades when OPR is extended further into shallower layers. The reason traces back to the suppression factor in Eq. 7, which relies on the segmentation head's weights trained on the deepest features: prototypes from shallow features carry weaker class semantics, so projecting them onto the segmentation head yields unreliable foreground assignments. Background prototypes are then mistaken for foreground and aggregated into the reconstruction, actively injecting the interference OPR is meant to remove.

A sensitivity analysis of λ_D, λ_B, and K on FLSMD with 2% labeled data is provided in the appendix (Figure A1); mIoU stays within 4 points of the optimum across all swept ranges, and each hyperparameter exhibits a clear single peak at the chosen default $\lambda_D = 0.1$, $\lambda_B = 0.05$, $K = 16$.

5 Conclusion

This paper identified a failure mode that emerges when imaging noise and long-tailed class distributions interact within the pseudo-labeling pipeline of semantic segmentation: a vicious cycle in which noise suppresses tail-class confidence, fixed thresholds turn that suppression into systematic exclusion, and the resulting absence of supervision lets tail features keep degrading. We argued that breaking this cycle requires intervention at both the feature level and the threshold level, and instantiated this principle as FTC-Seg, in which Orthogonal Prototype Reconstruction purifies pixel features against noise-induced entanglement and Adaptive Threshold Calibration replaces the fixed threshold with class-specific ones driven by an internal learning signal and an external distribution signal. Across four benchmarks spanning acoustic speckle, underwater optical scattering and adverse-weather imaging, FTC-Seg consistently outperforms state-of-the-art methods in overall mIoU, with the largest absolute gains concentrated on the rarest tail classes of each dataset, and the most substantial improvements in pseudo-label quality also occurring on tail classes.

Acknowledgments and Disclosure of Funding

We thank the Graduate School of Information, Production and Systems, Waseda University, for providing GPU computing resources. This work received no external financial support. The authors declare no competing interests.

References

[1] Changrui Chen, Jungong Han, and Kurt Debattista. Virtual category learning: A semi-supervised learning method for dense prediction with extremely limited labels. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5595–5611, 2024.

[2] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2613–2622, 2021.

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88 (2):303–338, 2010.

[5] Longhan Feng, Zihuang Cao, Ali Luo, Yuanhao Guo, Shuilian Yao, Yixin Guo, Qi Jia, and Yu Liu. Semi-supervised source detection in astronomical images: New benchmark and strong baseline. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 10901–10912, 2026.

[6] Ping Guo, Chengzhou Li, Guanchen Meng, Qi Jia, Jinyuan Liu, Zhu Liu, Yu Liu, Zhongxuan Luo, and Xin Fan. Ctfs: Collaborative teacher framework for forward-looking sonar image semantic segmentation with extremely limited labels. *arXiv preprint arXiv:2603.21071*, 2026.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Prantik Howlader, Srijan Das, Hieu Le, and Dimitris Samaras. Beyond pixels: Semi-supervised semantic segmentation with a multi-scale patch-based multi-label classifier. In *European Conference on Computer Vision*, pages 342–360. Springer, 2024.

[9] Lukas Hoyer, David Joseph Tan, Muhammad Ferjad Naeem, Luc Van Gool, and Federico Tombari. Semivl: Semi-supervised semantic segmentation with vision-language guidance. In *European Conference on Computer Vision*, pages 257–275. Springer, 2024.

[10] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021.

[11] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 1769–1776. IEEE, 2020.

[12] Chengzhou Li, Ping Guo, Guanchen Meng, Qi Jia, Jinyuan Liu, Zhu Liu, Xiaokang Liu, Yu Liu, Zhongxuan Luo, and Xin Fan. Rsod: Reliability-guided sonar image object detection with extremely limited labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6055–6063, 2026.

[13] Yu Liu, Longhan Feng, Qi Jia, Zezheng Liu, and Zi-Huang Cao. Two teachers are better than one: semi-supervised elliptical object detection by dual-teacher collaborative guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6355–6363, 2024.

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[15] Jaemin Na, Jung-Woo Ha, Hyung Jin Chang, Dongyoon Han, and Wonjun Hwang. Switching temporary teachers for semi-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 36:40367–40380, 2023.

[16] Youngtaek Oh, Dong-Jin Kim, and In So Kweon. Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9786–9796, 2022.

[17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[18] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10765–10775, 2021.

[19] Sebastian Scherer, Robin Schön, and Rainer Lienhart. Pseudo-label noise suppression techniques for semi-supervised semantic segmentation. *arXiv preprint arXiv:2210.10426*, 2022.

[20] Deepak Singh and Matias Valdenegro-Toro. The marine debris dataset for forward-looking sonar semantic segmentation. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 3741–3749, 2021.

[21] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[22] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[23] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022.

[24] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10857–10866, 2021.

[25] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7236–7246, 2023.

[26] Lihe Yang, Zhen Zhao, and Hengshuang Zhao. Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):3031–3048, 2025.

[27] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in neural information processing systems*, 34:18408–18419, 2021.

[28] HaiKuan Zhang, Haitao Li, Xiufeng Zhang, Guanyu Yang, Atao Li, Weisheng Du, Shanshan Xue, and Chi Liu. Noise-robust consistency regularization for semi-supervised semantic segmentation. *Neural Networks*, 184:107041, 2025.

[29] Zekai Zhang, Qinghui Chen, Maomao Xiong, Shijiao Ding, Zhanzhi Su, Xinjie Yao, Yiming Sun, Cong Bai, and Jinglin Zhang. Zero-shot learning in industrial scenarios: New large-scale benchmark, challenges and baseline. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10357–10366, 2025.

A Technical appendices and supplementary material

Table 7: Joint calibration across four datasets and three labeled-data ratios. M3 uses EMA+OPR, M4 uses EMA+ATC, and M5 uses EMA+OPR+ATC. Results are mean mIoU (%) over three matched seeds shared by M3–M5 within each setting; these seeds differ from those used in the main comparison tables. $\Delta = \mathrm{M5} - \max(\mathrm{M3}, \mathrm{M4})$ is measured in percentage points. Two-sided paired t-tests compare M5 against the stronger individual module selected by mean mIoU. Reported p-values are Holm-adjusted across the 12 settings; † denotes $p \geq 0.05$.

Dataset	Labels	M3	M4	M5	Δ	Holm p
FLSMD	2%	59.80	60.17	**63.07**	+2.90	0.0436
	5%	66.72	66.49	**68.84**	+2.12	0.0225
	10%	71.24	70.99	**73.49**	+2.25	0.0356
FSSG	2%	60.87	61.24	**63.20**	+1.96	0.0448
	5%	62.73	62.46	**66.31**	+3.58	0.0160
	10%	63.97	64.16	**68.01**	+3.85	0.0086
SUIM	2%	64.49	63.79	**67.04**	+2.55	0.0192
	5%	67.14	67.33	**70.33**	+3.00	0.0041
	10%	67.53	68.17	**70.49**	+2.32	0.0086
ACDC	2%	58.32	59.07	**61.03**	+1.96	$0.0769^{\dagger}$
	5%	64.92	64.12	**68.29**	+3.37	0.0448
	10%	67.37	68.07	**71.60**	+3.53	0.0356

Table 8: Prototype formulation comparison on FSSG with 2% labels using the same seed. Original OPR uses 16 learnable prototypes without predefined class identities. Class-anchored OPR uses one prototype per class, initialized from labeled class-mean features and assigned a fixed class identity. Tail mIoU measures segmentation on rare classes; pseudo-label mIoU evaluates accepted pseudo-labels against ground truth on unlabeled training pixels. All metrics are percentages.

Method	mIoU	Tail mIoU	Pseudo-label mIoU
EMA baseline	58.78	5.33	53.49
Original OPR	**62.55**	**17.19**	**73.86**
Class-anchored OPR	54.87	8.93	58.25

Table 9: Sensitivity to misspecifying the foreground set $\mathcal{C}_{fg}$ on FSSG with 2% labels. All variants use EMA+OPR without ATC, a fixed pseudo-label threshold of 0.95, and the same seed. Background and Diver have category IDs 0 and 10, respectively. Only the gating set changes; segmentation output classes remain unchanged. IoU values are percentages, and Δ denotes the mIoU change from the correct setting in percentage points.

Setting	$\mathcal{C}_{fg}$	mIoU	Δ	Diver IoU
Correct	$\{1, \ldots, 10\}$	62.55	–	17.50
Add background	$\{0, \ldots, 10\}$	59.80	-2.75	10.11
Omit Diver	$\{1, \ldots, 9\}$	58.42	-4.13	11.49
Add background + omit Diver	$\{0, \ldots, 9\}$	53.09	-9.46	2.47

Table 10: Sensitivity of the noise–long-tail interaction to corruption severity. Each cell reports the corruption parameter and interaction $I = CL - OL - CB + OB$ in percentage points. OB, OL, CB, and CL denote tail-class pseudo-label filtering rates under original-balanced, original-long-tail, corrupted-balanced, and corrupted-long-tail conditions, respectively. Positive I indicates super-additive exclusion. Larger speckle σ, smaller scattering transmission t, and larger a with smaller b for fog indicate stronger corruption. The checkpoint, image lists, class definitions, threshold, and random seed are held fixed across severity levels.

Dataset / corruption	Mild: parameter / I	Medium: parameter / I	Severe: parameter / I
FLSMD / speckle	$\sigma = 0.15$ / $+0.89$	$\sigma = 0.35$ / $+1.83$	$\sigma = 0.60$ / $+2.92$
SUIM / scattering	$t = 0.80$ / $+1.04$	$t = 0.60$ / $+1.56$	$t = 0.40$ / $+1.22$
ACDC / fog	$(a, b) = (1.5, 2.0)$ / -0.11	$(a, b) = (2.5, 1.7)$ / $+1.55$	$(a, b) = (3.0, 1.4)$ / $+2.23$

Table 11: Single-seed controlled evaluation on Pascal VOC using the public 1/16 semi-supervised split with 155 labeled and 5,507 unlabeled training images. A long-tailed distribution with imbalance ratio 20 is constructed, and Gaussian noise with $\sigma = 0.18$ is applied to all labeled and unlabeled training images in the noisy condition. Both conditions use the same training split and unchanged official validation set. Results are validation mIoU (%). Noise-induced change is the noisy result minus the long-tail-only result; gains and changes are reported in percentage points.

Method	Long-tail only	Long-tail + noise	Noise-induced change
Baseline	70.76	63.18	-7.58
FTC-Seg	**75.36**	**69.11**	-6.25
FTC-Seg gain	$+4.60$	$+5.93$	–

Table 12: Inference cost of compared methods and individual modules with DINOv2-S and a DPT head. Parameters and FLOPs are measured for a single $1 \times 3 \times 518 \times 518$ input. FPS is measured after warm-up on an RTX 4090. The upper block reports method comparisons; the lower block reports component ablations. FSSG results with 2% labels are included as performance references. ATC operates during training and adds no inference-time computation.

Method	Params (M)	FLOPs (G)	FPS	mIoU (%)
Method comparison				
Labeled only	24.79	56.94	122.62	35.61
UniMatch V2	24.79	56.94	122.62	58.78
FTC-Seg	24.81	56.96	120.93	63.20
Component ablation				
M2: EMA	24.79	56.88	122.62	58.75
M3: EMA + OPR	24.81	56.92	120.93	62.55
M4: EMA + ATC	24.79	56.94	122.62	61.90
M5: EMA + OPR + ATC	24.81	56.96	120.93	63.20

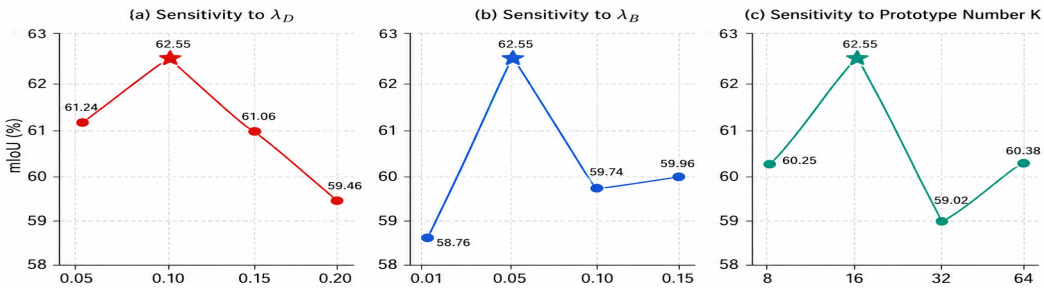


Figure A1: Sensitivity analysis of key hyperparameters in FTC-Seg framework on the FLSMD dataset with 2% labeled data.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 claim that imaging noise and long-tailed distributions jointly produce a vicious cycle of pseudo-label degradation that requires dual-calibration at both the feature and threshold levels. Section 3 instantiates this principle as FTC-Seg and Section 4 verifies it across four benchmarks spanning three noise modalities.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: OPR identifies foreground prototypes through the segmentation head's class weights and therefore depends on the head producing meaningful class decisions, which may degrade under extreme label scarcity where head classes are themselves poorly learned. Our empirical evaluation also covers only three visual imaging-noise modalities and does not verify whether the same dual-calibration principle transfers to label-noise or non-visual signals.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results such as theorems or formal proofs. All claims are supported empirically through the experiments in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 fully specifies the architecture and training objectives of FTC-Seg with all formulas numbered and cross-referenced, and Section 4 reports the datasets, encoders, optimizer, learning rate, training epochs, hardware, and labeled-data ratios needed to reproduce the main experiments. All comparison methods are reproduced from publicly available code under identical settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code is not released at submission time to preserve anonymity but will be made publicly available upon acceptance. All datasets used are publicly available and cited in Section 4, and the implementation details in Section 4 are sufficient to reproduce the main results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 specifies the four datasets, the labeled-data ratios, the encoder architectures, the optimizer with its learning rate and weight decay, the training schedule, and the hardware used. The hyperparameters of OPR and ATC are defined in Section 3 and their selected values are reported in the hyperparameter sensitivity analysis.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Tables 1 and 2 report standard deviations averaged over 3 random seeds for each method-dataset-ratio configuration, with seed initialization being the source of variability. The reported values are 1-sigma standard deviations.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section 4 reports that all experiments are conducted on a single NVIDIA RTX 3090 GPU and trained for 240 epochs. Table 5(a) further reports peak training memory and per-image inference time for OPR under different configurations.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research uses only publicly available datasets and does not involve human subjects, sensitive data, or any ethical concerns. Anonymity is preserved throughout the submission.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The proposed method improves segmentation in noisy and long-tailed scenarios with positive applications in marine monitoring and autonomous driving, while potential negative uses (e.g., surveillance under degraded imaging) and residual errors on safety-critical tail classes warrant cautious deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any pre-trained generative models or scraped datasets that pose a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets (FLSMD, FSSG, SUIM, ACDC) and baseline methods used in this paper are publicly available and properly cited in Section 4.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new datasets or pre-trained models are released and the code will be made publicly available upon acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology of this paper does not involve LLMs as any important or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.