A BENCHMARK FOR PRESENCE-UNCERTAIN FEW-SHOT SEGMENTATION IN REAL-WORLD REMOTE SENSING SCENARIOS

Youngtack Oh[1*] **Jeongwon Ryu**[1*] **Minseok Seo**[2†]

[1]SI Analytics
[2]Korea Advanced Institute of Science and Technology (KAIST)

{ytoh96,rjw0926}@si-analytics.ai
minseok.seo@kaist.ac.kr

ABSTRACT

Few-shot segmentation (FSS) is typically evaluated under a positive-class assumption, where the reference class is guaranteed to appear in every target image. This assumption is particularly fragile in remote sensing imagery, where large-scale scenes are processed in a patch-wise manner and target objects are often entirely absent. To better reflect real-world application scenarios, evaluation must consider not only where to segment, but also whether segmentation should be performed at all. To address this gap, we introduce the **Presence-Uncertain Few-Shot Segmentation (PU-FSS) Benchmark**, centered on the aerial imagery dataset iSAID, which naturally exhibits presence uncertainty due to large-scale scenes. We further include standard natural image datasets to verify that the proposed evaluation protocol is not limited to a single domain. We also propose a lightweight **training-free prototype-debiasing (PD) module** that suppresses background bias in similarity computation, enabling reliable abstention when the reference class is absent. The proposed PU-FSS provides an evaluation protocol aligned with the requirements of large-scale remote sensing image analysis. Our results show that the high performance of training-free FSS models under the positive-class assumption does not translate to reliability in real-world settings, and that prototype debiasing provides a practical means to evaluate such models under presence uncertainty, particularly in remote sensing imagery. Our code and dataset are available at PUFSS.

1 INTRODUCTION

Few-shot segmentation (FSS) aims to segment novel classes using only a few annotated reference images, and has recently attracted attention as a practical alternative to dense annotation. In remote sensing imagery, however, the assumptions underlying standard FSS evaluation are particularly fragile. Specifically, most benchmarks rely on the *positive-class assumption*, where the reference class is guaranteed to appear in every target image.

Such imagery typically covers large-scale scenes and is processed in a patch-wise manner during inference, resulting in a high proportion of patches where the target object is entirely absent. In such settings, models must decide not only *where* to segment but also *whether* to segment. Under the positive-class assumption of current FSS benchmarks, training-free generalist models (Zhang et al., 2023; Liu et al., 2023b; Zhang et al., 2024) typically select top-k candidate regions based on similarity and apply SAM-based refinement. This design yields strong benchmark results but leads to unreliable predictions in presence-absent scenarios (Figure 1). While several studies (Kang & Cho, 2022; Liu et al., 2023a; Yu & Li, 2024) have questioned the positive-class assumption, such analyses remain largely unexplored in remote sensing, where a standardized, cross-domain evaluation protocol is still missing. To address this gap, we introduce the **Presence-Uncertain Few-Shot Segmentation (PU-FSS) Benchmark**, which explicitly incorporates both target-present and target-absent scenarios. The benchmark enables systematic evaluation of not only segmentation

[*]Equal contribution
[†]Corresponding author

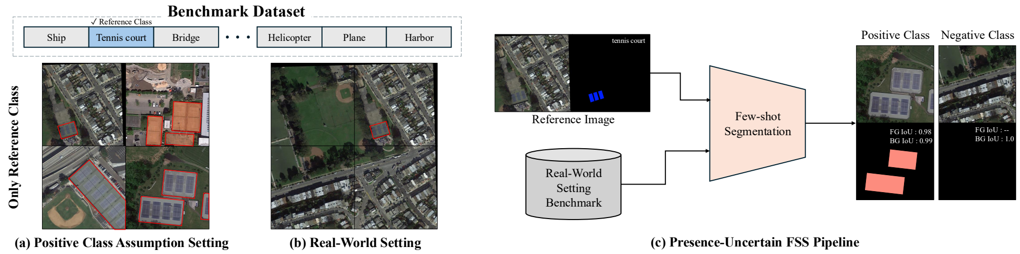


Figure 1: Overview of the proposed Presence-Uncertain Few-shot Segmentation (PU-FSS) Benchmark. (a) *Positive Class Assumption* setting, where all evaluation samples contain the reference class. (b) *Real-world* setting, where evaluation includes both reference-matching and non-matching samples (red polygons = positive; no polygon = negative). (c) Few-shot segmentation pipeline evaluated under the PU-FSS protocol.

accuracy when the target class is present, but also abstention reliability under presence uncertainty, reflecting practical remote sensing settings. To assess generalization, PU-FSS spans both aerial and natural image domains, and reorganizes existing datasets into presence-aware evaluation splits (Appendix A).

The proposed PU-FSS provides an evaluation protocol aligned with real-world large-scale image analysis. Extensive experiments reveal that existing FSS methods exhibit critical limitations under presence uncertainty, motivating both the benchmark and our prototype debiasing module.

2 PRESENCE-UNCERTAIN FSS (PU-FSS) BENCHMARK

2.1 PRELIMINARIES

In standard few-shot segmentation (FSS) benchmarks, models are tasked with predicting a pixel-wise mask of a reference class in a target image, given a small set of annotated reference images. Typically, a dataset with C classes is partitioned into N disjoint groups to define base and novel classes. During evaluation, one group is treated as novel classes, while the remaining groups define the base classes used for model development. Each evaluation instance consists of a target image x^c_{tar} together with a reference set $R^c = \{(x^c_{\text{ref},k}, m^c_{\text{ref},k})\}_{k=1}^{K}$, where $m^c_{\text{ref},k}$ denotes a binary mask for class c.

Performance is typically measured by the Intersection-over-Union (IoU) between $\tilde{y}$ and the ground-truth mask y^c. Crucially, this evaluation protocol relies on the *positive-class assumption*, which guarantees that the reference class is always present in the target image. As a result, models are never required to abstain from prediction, even when the reference class is absent in practice.

2.2 DETAILS OF THE PU-FSS BENCHMARK

The PU-FSS benchmark is centered on the aerial imagery dataset iSAID (15 classes), where large-scale scenes, small objects, and densely packed instances naturally induce presence uncertainty. To evaluate generalization capability, we further incorporate three widely used natural image datasets: PASCAL VOC (20 classes), PASCAL-Context (59 classes), and COCO-Stuff (171 classes). While PASCAL VOC provides clean object-centric annotations, PASCAL-Context and COCO-Stuff cover diverse object and stuff categories, enabling evaluation across a broader range of object and stuff categories.

Entropy-based uncertainty. To investigate how segmentation performance varies across different levels of input difficulty, we estimate the uncertainty of each evaluation image using the predictive entropy of a pretrained segmentation model.

Given softmax probabilities $q(y_i \mid x)$ at pixel i for class $y_i \in \{1, \ldots, C\}$, the pixel-level entropy is defined as

$$H_i(x) = -\sum_{c=1}^{C} q(y_i = c \mid x) \log q(y_i = c \mid x), \tag{1}$$

Methods	iSAID-PU			Pascal-Context-PU			COCO-Stuff-PU			Pascal-PU		
	Pos.-easy	Pos.-hard	Negative	Pos.-easy	Pos.-hard	Negative	Pos.-easy	Pos.-hard	Negative	Pos.-easy	Pos.-hard	Negative
Trained generalists												
SegIC(Meng et al., 2024)	27.07	15.13	**99.62** / **56.84**	43.38	33.18	**94.99** / 30.13	34.11	25.25	**95.93** / 34.27	64.77	53.35	**94.45** / 46.47
SegGPT(Wang et al., 2023a)	24.96	6.55	99.21 / 44.04	47.70	31.61	94.08 / **58.02**	44.90	24.91	95.54 / 52.25	**79.16**	60.08	93.62 / **55.88**
Training-free generalists												
PerSAM(Zhang et al., 2023)	8.47	4.26	93.49 / —	34.97	20.56	78.44 / —	29.97	16.77	86.30 / —	51.25	31.92	85.20 / —
Matcher(Liu et al., 2023b)	25.08	10.82	88.53 / —	48.11	30.23	79.46 / —	47.52	29.41	82.27 / —	60.42	42.58	82.11 / —
GF-SAM(Zhang et al., 2024)	30.42	12.40	89.76 / —	**56.01**	**39.97**	81.79 / —	54.34	37.88	83.95 / —	73.17	**65.67**	84.98 / —
Ours	**45.11**	**25.78**	92.05 / 33.40	55.21	39.02	81.95 / 33.80	**61.23**	**48.41**	88.47 / **59.10**	72.81	65.30	85.02 / 28.40

Table 1: One-shot segmentation performance on the PU-FSS benchmark. For positive classes, we report Foreground IoU (%). For negative classes, we report Background IoU and Scene-level Accuracy (%). All results are averaged over 5-fold cross-validation.

and the image-level entropy for class c is computed by averaging over pixels belonging to its reference mask m^c:

$$H(x^c) = \frac{1}{|m^c|} \sum_{i \in m^c} H_i(x). \tag{2}$$

This entropy score serves as a proxy for sample difficulty, under the assumption that higher predictive uncertainty reflects intrinsic dataset ambiguity rather than model-specific noise (Smith et al., 2014; Kwon et al., 2024). Empirically, samples with ambiguous visual content (e.g., small objects or fuzzy boundaries) yield consistently high entropy across architectures, making the relative ordering of difficulty largely model-invariant (Cui et al., 2023; Geirhos et al., 2020; Toneva et al., 2018).

For practical estimation, we adopt PSPNet (Zhao et al., 2017) from MMSegmentation (Contributors, 2020) as the entropy estimator. Because the entropy captures dataset-driven ambiguity, the resulting partition into *positive-easy* and *positive-hard* subsets transfers robustly to other segmentation models, enabling standardized and reproducible evaluation.

Reference (Visual Prompts). From each dataset, we select reference samples exclusively from the *positive-easy* subset to ensure clear and unambiguous visual prompts. Each reference consists of an image paired with its corresponding target-class mask. To mitigate reference-image bias, we provide **five** distinct reference samples for every source corpus.

3 Methodology

Motivation. We propose a lightweight and training-free **Prototype-Debiasing (PD)** module for few-shot segmentation under presence uncertainty. The core challenge in this setting is that similarity-based generalist models inevitably produce non-zero responses by matching global scene or class-level semantics, even when the reference object is absent. PD addresses this issue by explicitly normalizing similarity computation to focus on object-specific cues rather than global context. As a result, similarity responses naturally collapse in no-target scenarios, enabling reliable abstention without heuristic thresholds. The proposed method consists of three stages: (i) debiased prototype construction from the reference image, (ii) orthogonal similarity estimation on the target image, and (iii) prompt-based segmentation using the resulting similarity map. Detailed mathematical formulations are provided in Appendix B.

Prototype-Debiased Similarity and Segmentation. Given a reference image and its binary mask, we extract spatial features and a global class token from a frozen Vision Transformer. A foreground prototype is obtained by aggregating masked features. however, it is often biased toward global scene and background semantics. PD removes this bias by orthogonally projecting out class-level and background components, yielding an object-specific prototype. For a target image, features are processed with the same encoder and similarly decorrelated from the global token, suppressing spurious similarity induced by global context. Cosine similarity between the debiased prototype and target features produces an orthogonal similarity map that responds only to true object alignment. This map is used to derive point prompts for segmentation: when the reference object is absent, similarity responses remain low and no prompts are selected, without heuristic thresholds. Selected prompts are passed to the Segment Anything Model (Kirillov et al., 2023), and the final mask is obtained using GF-SAM-style graph filtering.

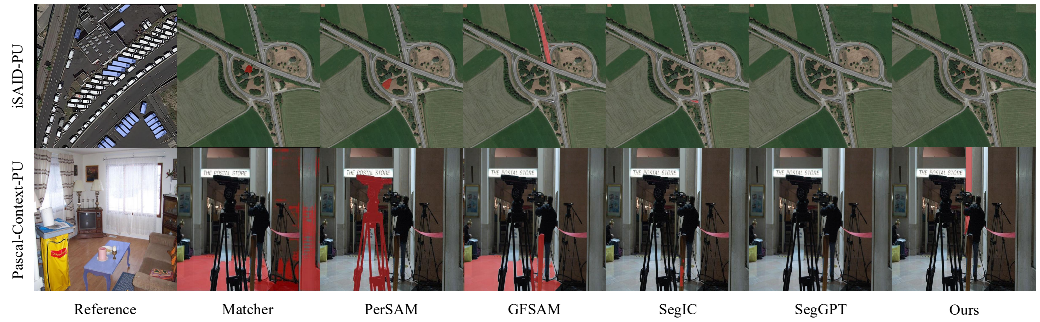


Figure 2: Qualitative comparison on the negative set of PU-FSS. Training-free methods such as Matcher, PerSAM and GF-SAM produce clear false positives, while trained generalist models (SegIC and SegGPT) avoid such errors.

4 Experimental Results

Datasets. We primarily evaluate our approach on the proposed **PU-FSS Benchmark**, which explicitly includes both positive and negative cases where the reference class may or may not appear in the target image. PU-FSS is constructed from iSAID, PASCAL VOC, PASCAL-Context, and COCO-Stuff, with iSAID images tiled into non-overlapping 896×896 patches due to their large resolution. The benchmark is organized into five folds to mitigate reference-image bias. In addition, we report results on standard few-shot segmentation benchmarks (iSAID-5^i, PASCAL-5^i, and COCO-20^i) to verify that our method remains competitive under conventional evaluation protocols. Further details are provided in Appendix D.

Results on the PU-FSS Benchmark. Table 1 summarizes the quantitative results on the PU-FSS Benchmark. Overall, performance on **iSAID-PU** is consistently lower than on natural-image datasets, reflecting the intrinsic difficulty of remote sensing imagery, where objects are small, densely distributed, and embedded in cluttered backgrounds.

Even under the positive-class assumption, training-free generalist models exhibit a pronounced performance gap between the positive-easy and positive-hard subsets on iSAID-PU, highlighting the challenges posed by small, ambiguous objects embedded in cluttered remote sensing scenes. This gap further widens as evaluation moves to harder subsets.

More critically, PU-FSS exposes a fundamental limitation of training-free approaches. In negative cases, where the reference class is absent, similarity-based methods inevitably generate false positives, making reliable scene-level evaluation impossible. Figure 2 qualitatively illustrates this behavior in remote sensing imagery, with additional qualitative results and visualizations of the proposed prototype debiasing strategy provided in Appendix E.

Although our method does not eliminate all false positives, it substantially reduces them and, importantly, avoids forcing a foreground prediction when the reference class is absent. This enables negative scenes to be explicitly observed and evaluated, underscoring the necessity of presence-uncertain benchmarks such as PU-FSS for remote sensing applications. Results on standard few-shot segmentation benchmarks are reported in Appendix F, and further ablation studies are provided in Appendix G.

5 Conclusion

In this work, we introduced the **Presence-Uncertain Few-Shot Segmentation (PU-FSS) Benchmark** to study few-shot segmentation under realistic conditions where the target object may be absent, with a particular focus on remote sensing imagery. Our experiments reveal that training-free few-shot methods, which perform well under the positive-class assumption, are especially vulnerable in remote sensing scenarios with small objects and cluttered scenes. By enabling explicit evaluation of such failure cases, PU-FSS provides a practical framework for analyzing and comparing few-shot segmentation methods in real-world remote sensing imagery.

REFERENCES

MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020.

Peng Cui, Dan Zhang, Zhijie Deng, Yinpeng Dong, and Jun Zhu. Learning sample difficulty from pre-trained models for reliable prediction. *Advances in Neural Information Processing Systems*, 36:25390–25408, 2023.

Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in neural information processing systems*, 33:13890–13902, 2020.

Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, pp. 108–126. Springer, 2022.

Dahyun Kang and Minsu Cho. Integrative few-shot learning for classification and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9979–9990, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Hyunjin Kwon, Matthew Greenberg, Colin Bruce Josephson, and Joon Lee. Measuring the prediction difficulty of individual cases in a dataset using machine learning. *Scientific Reports*, 14(1): 10474, 2024.

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23592–23601, 2023a.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310*, 2023b.

Lingchen Meng, Shiyi Lan, Hengduo Li, Jose M Alvarez, Zuxuan Wu, and Yu-Gang Jiang. Segic: Unleashing the emergent correspondence for in-context segmentation. In *European Conference on Computer Vision*, pp. 203–220. Springer, 2024.

Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6941–6952, 2021.

Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 622–631, 2019.

Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23641–23651, 2023.

Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017.

Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier. An instance level analysis of data complexity. *Machine learning*, 95(2):225–256, 2014.

Yanpeng Sun, Jiahui Chen, Shan Zhang, Xinyu Zhang, Qiang Chen, Gang Zhang, Errui Ding, Jingdong Wang, and Zechao Li. Vrp-sam: Sam with visual reference prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23565–23574, 2024.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

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

Yuan Wang, Naisong Luo, and Tianzhu Zhang. Focus on query: Adversarial mining transformer for few-shot segmentation. *Advances in neural information processing systems*, 36:31524–31542, 2023b.

Qianxiong Xu, Xuanyi Liu, Lanyun Zhu, Guosheng Lin, Cheng Long, Ziyue Li, and Rui Zhao. Hybrid mamba for few-shot segmentation. *Advances in Neural Information Processing Systems*, 37:73858–73883, 2024.

Xiwen Yao, Qinglong Cao, Xiaoxu Feng, Gong Cheng, and Junwei Han. Scale-aware detailed matching for few-shot aerial image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021.

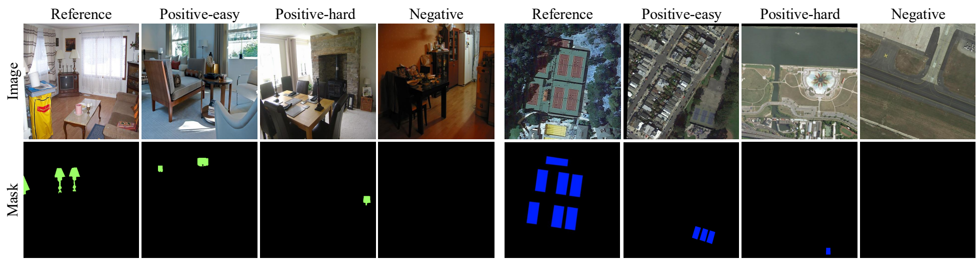
Zhihan Yu and Ruifan Li. Revisiting counterfactual problems in referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13438–13448, 2024.

Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. *Advances in Neural Information Processing Systems*, 37: 33232–33261, 2024.

Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

A Example of the iSAID-PU dataset and Pascal-PU dataset



(a) Pascal-PU dataset (b) iSAID-PU dataset

Figure Appendix.1: Example of the **Pascal-PU** dataset for the class *light* (a) and the **iSAID-PU** dataset for the class *tennis_court* (b). From left to right we show the visual prompt (*Reference*) and representative reference images from the *positive-easy*, *positive-hard*, and *negative* subsets, which are defined based on entropy-based uncertainty.

B Mathematical Details

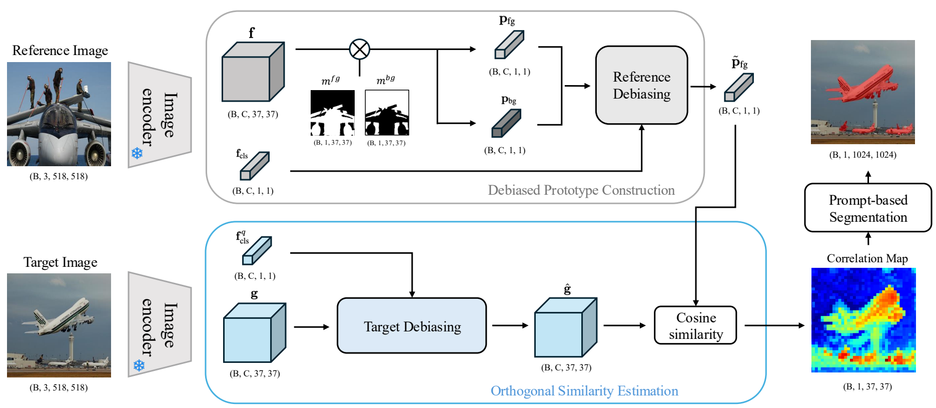


Figure Appendix.2: Prototype-Debiasing (PD) module overview: (i) extract and orthogonally purify the foreground prototype from a reference; (ii) compute an orthogonal cosine map against decorrelated reference features; (iii) perform prompt-based SAM segmentation.

Given a reference image with a binary mask, we extract spatial features $\{\mathbf{f}_i\}_{i=1}^{N}$ and a global class token $\mathbf{f}_{\text{cls}}$ from a frozen Vision Transformer encoder, where $N = H \times W$. The binary mask is flattened into $\mathbf{m} \in \{0,1\}^N$, with $m_i = 1$ indicating foreground pixels.

Foreground and background prototypes are computed as

$$\mathbf{p}_{\text{fg}} = \frac{1}{\sum_i m_i} \sum_i m_i \mathbf{f}_i, \quad \mathbf{p}_{\text{bg}} = \frac{1}{\sum_i (1-m_i)} \sum_i (1-m_i)\mathbf{f}_i. \tag{3}$$

The global class token encodes scene-level semantics, while the background prototype captures dominant non-target information. To suppress these biases, we construct an orthonormal basis $\{\mathbf{u}_{\text{cls}}, \mathbf{u}_{\text{bg}}\}$ via Gram–Schmidt orthogonalization.

The debiased foreground prototype is defined as

$$\tilde{\mathbf{p}}_{\text{fg}} = \mathbf{p}_{\text{fg}} - \alpha(\mathbf{p}_{\text{fg}} \cdot \mathbf{u}_{\text{cls}})\mathbf{u}_{\text{cls}} - \beta(\mathbf{p}_{\text{fg}} \cdot \mathbf{u}_{\text{bg}})\mathbf{u}_{\text{bg}}, \tag{4}$$

where α and β control the projection strength. We set $\alpha = \beta = 0.6$ in all experiments.

This operation removes global and background components from the prototype, emphasizing features distinctive to the reference object.

B.1 Query-side Debiasing and Similarity Computation

For a target image, spatial features $\{\mathbf{g}_i\}_{i=1}^{N}$ and a global class token $\mathbf{f}^q_{\text{cls}}$ are extracted using the same encoder. Each spatial feature is orthogonally projected away from the normalized class token:

$$\hat{\mathbf{g}}_i = \mathbf{g}_i - \alpha_q(\mathbf{g}_i \cdot \mathbf{u}^q_{\text{cls}})\mathbf{u}^q_{\text{cls}}, \tag{5}$$

where $\mathbf{u}^q_{\text{cls}} = \mathbf{f}^q_{\text{cls}}/\|\mathbf{f}^q_{\text{cls}}\|$ and $\alpha_q = 0.6$.

Cosine similarity between the debiased prototype and debiased target features yields

$$s_i = \frac{\tilde{\mathbf{p}}_{\text{fg}} \cdot \hat{\mathbf{g}}_i}{\|\tilde{\mathbf{p}}_{\text{fg}}\|\|\hat{\mathbf{g}}_i\|}. \tag{6}$$

Reshaping $\{s_i\}$ produces the orthogonal cosine similarity map used for prompt selection.

C Implementation Details

Backbones and resolution. For fair comparison with prior work, we use a frozen DINOv2 ViT-L/14 encoder to extract spatial and class tokens (layer 23) and a ViT-H-based SAM for prompt-based mask generation. All input images are resized to 518×518 for feature encoding and 1024×1024 for SAM inference following prior work. All experiments were conducted on a single NVIDIA Quadro RTX 8000 GPU.

Projection strength. We explored the projection strength λ from 0.1 to 0.9. Small values retain noticeable feature bias, while large values over-suppress class-discriminative signals. We use $\lambda = 0.6$ as a stable default across datasets and subsets.

D Additional Dataset Details

PU-FSS benchmark construction. The proposed PU-FSS benchmark is constructed from four segmentation datasets: PASCAL VOC, PASCAL-Context, COCO-Stuff, and iSAID. For each dataset, evaluation episodes include both *positive* samples, where the reference class is present in the target image, and *negative* samples, where the reference class is absent. We refer to the resulting presence-uncertain variants as **Pascal-PU**, **Pascal-Context-PU**, **COCO-Stuff-PU**, and **iSAID-PU**, respectively.

Folds and reference selection. To mitigate reference-image bias, each PU-FSS variant is organized into five folds. For each fold, a fixed set of reference images is selected from the positive-easy subset, and all methods are evaluated using the same references. Results reported in the paper are averaged over the five folds.

iSAID tiling strategy. Due to the large spatial resolution of iSAID images, each image is divided into non-overlapping 896×896 tiles prior to evaluation. This tiling strategy follows common practice in remote sensing segmentation and ensures that both positive and negative patches naturally arise within a single scene. All other datasets are evaluated using their original image resolutions.

Standard FSS benchmarks. In addition to PU-FSS, we evaluate all methods on standard few-shot segmentation benchmarks, including PASCAL-5^i, COCO-20^i, and iSAID-5^i. These benchmarks follow the conventional positive-class assumption and are used to verify that the proposed prototype-debiasing module remains competitive under established evaluation protocols.

E Additional Qualitative Results

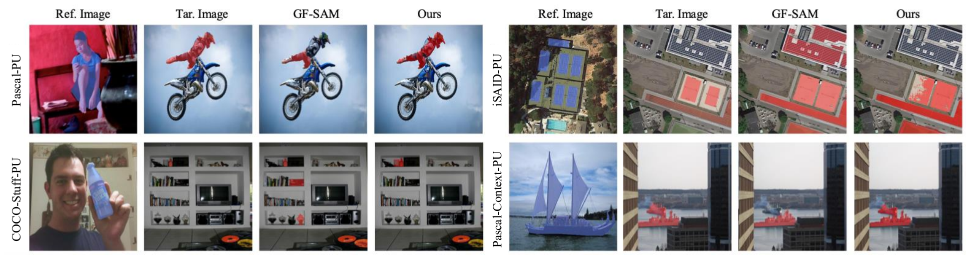


Figure Appendix.3: Qualitative comparison of our method with GF-SAM on Pascal-PU, COCO-Stuff-PU, iSAID-PU, and Pascal-Context-PU datasets. Blue regions on reference images indicate masks, while red regions on target images represent ground truth (GT) masks.

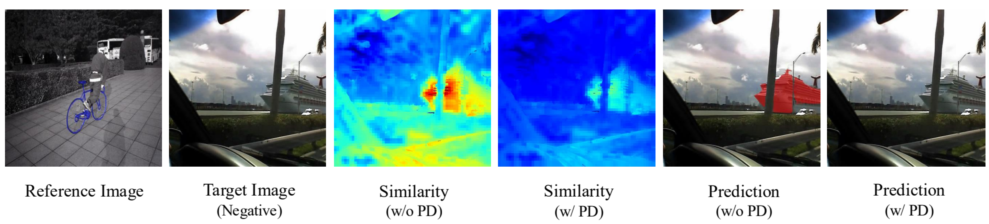


Figure Appendix.4: Visualization of similarity maps and predictions demonstrating the effectiveness of our Prototype Debiasing (PD) strategy.

Figure Appendix.3 provides qualitative comparisons on PU-FSS across datasets. Figure Appendix.4 visualizes similarity maps and predictions to illustrate the effect of PD. Figure 2 further compares methods on the negative subset.

F Additional Quantitative Results

Standard FSS benchmarks. Table Appendix.1 reports results on PASCAL-5^i, COCO-20^i, and iSAID-5^i for direct comparison with prior work.

Methods	Pascal-5^i (Shaban et al., 2017)	COCO-20^i (Nguyen & Todorovic, 2019)	iSAID-5^i (Yao et al., 2021)
specialist model			
HSNet(Min et al., 2021)$_{\text{CVPR21}}$	64.0	39.2	34.1
VAT(Hong et al., 2022)$_{\text{ECCV22}}$	65.5	41.3	–
HDMNet(Peng et al., 2023)$_{\text{CVPR23}}$	69.4	50.0	–
AMFormer(Wang et al., 2023b)$_{\text{NeurIPS23}}$	70.1	51.0	–
HMNet(Xu et al., 2024)$_{\text{NeurIPS24}}$	70.4	52.1	–
generalist model			
PerSAM(Zhang et al., 2023)$_{\text{ICLR24}}$	43.1	23.0	19.2
PerSAM-F(Zhang et al., 2023)$_{\text{ICLR24}}$	48.5	23.5	20.3
Matcher(Liu et al., 2023b)$_{\text{ICLR24}}$	68.1	52.7	33.3
VRP-SAM(Sun et al., 2024)$_{\text{CVPR24}}$	71.9	53.9	–
GF-SAM(Zhang et al., 2024)$_{\text{NeurIPS24}}$	72.1	58.7	47.1
Ours	**72.4**(0.3)	**60.1**(1.4)	**53.7**(6.6)

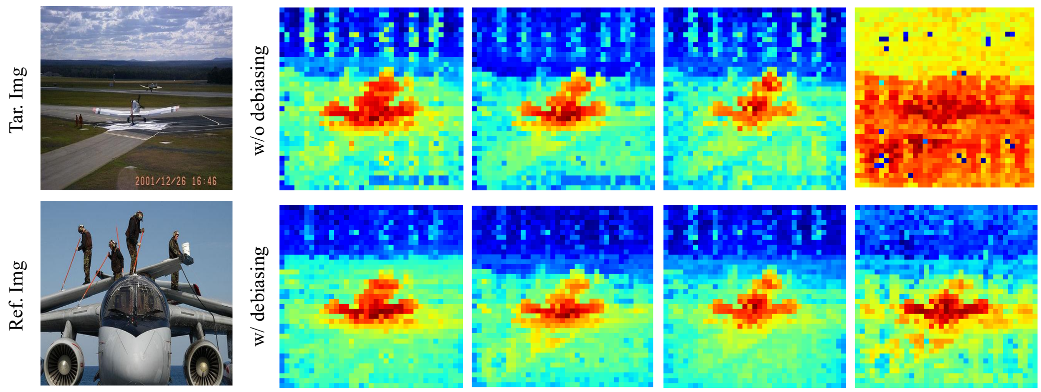
Table Appendix.1: Mean IoU performance on standard FSS benchmarks (Pascal-5^i, COCO-20^i, and iSAID-5^i), where mIoU is the average of foreground IoU and background IoU.

G Additional Ablation Studies

Ablation. Table Appendix.2 reports the effect of PD when integrated into baselines.

Variant	Pascal-PU		
	Pos.–easy	Pos.–hard	Negative
Baseline (Matcher(Liu et al., 2023b))	69.2	44.7	64.3
+ Prototype Debiasing	73.2(4.0)	46.7(2.0)	66.6(2.3)
Baseline (GF-SAM(Zhang et al., 2024))	75.2	55.8	89.4
+ Prototype Debiasing	76.5(1.3)	58.1(2.3)	90.1(0.7)

Table Appendix.2: Effect of the PD module on multiple baselines (**Pascal-PU**.



(a) Inputs (b) ViT-S (c) ViT-B (d) ViT-L (e) ViT-G

Figure Appendix.5: **Qualitative ablation study of Prototype Debiasing.** The leftmost column shows the target (top) and reference (bottom) images, while the remaining columns depict similarity maps produced by DINOv2 with four Vision Transformer backbones—*(b)* ViT-S, *(c)* ViT-B, *(d)* ViT-L, and *(e)* ViT-G—on the same query.

Figure Appendix.5 shows qualitative ablations of similarity maps across ViT backbones, demonstrating that PD consistently mitigates feature bias across model scales.