Bridge the Points: Graph-based Few-shot Segment Anything Semantically

Anqi Zhang[1], Guangyu Gao[1]*, Jianbo Jiao[2], Chi Harold Liu[1], and Yunchao Wei[3]
[1]School of Computer Science, Beijing Institute of Technology
[2]The MIx group, School of Computer Science, University of Birmingham
[3]WEI Lab, Institute of Information Science, Beijing Jiaotong University
andy_zaq@outlook.com

Abstract

The recent advancements in large-scale pre-training techniques have significantly enhanced the capabilities of vision foundation models, notably the Segment Anything Model (SAM), which can generate precise masks based on point and box prompts. Recent studies extend SAM to Few-shot Semantic Segmentation (FSS), focusing on prompt generation for SAM-based automatic semantic segmentation. However, these methods struggle with selecting suitable prompts, require specific hyperparameter settings for different scenarios, and experience prolonged one-shot inference time due to the overuse of SAM, resulting in low efficiency and limited automation ability. To address these issues, we propose a simple yet effective approach based on graph analysis. In particular, a Positive-Negative Alignment module dynamically selects the point prompts for generating masks, especially uncovering the potential of the background context as the negative reference. Another subsequent Point-Mask Clustering module aligns the granularity of masks and selected points as a directed graph, based on mask coverage over points. These points are then aggregated by decomposing the weakly connected components of the directed graph in an efficient manner, constructing distinct natural clusters. Finally, the positive and overshooting gating, benefiting from graph-based granularity alignment, aggregate high-confident masks and filter out the false-positive masks for final prediction, without relying on additional hyperparameters and redundant mask generation. Extensive experimental analysis across tasks including the standard FSS, One-shot Part Segmentation, and Cross Domain FSS validate the effectiveness and efficiency of the proposed approach, surpassing state-of-the-art generalist models with a mIoU of 58.7% on COCO-20^i and 35.2% on LVIS-92^i. The project page of this work is: https://andyzaq.github.io/GF-SAM/.

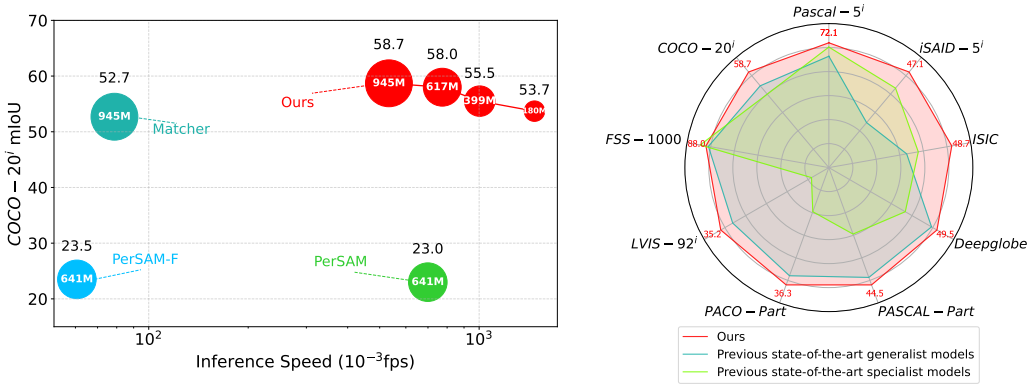
1 Introduction

Previous semantic segmentation methods [1–8], which rely on the pixel-level classification, often struggle with generalization and overfitting due to limited labeled data. In addition, recent approaches, such as MaskFormer [9], have shifted the paradigm to mask-based classification, offering a more flexible approach to improving the segmentation performance by exploiting the consistency and completeness of generated class-agnostic masks. The Segment Anything Model (SAM) [10] further marks a significant advancement by utilizing extensive pre-training on huge-scale dataset SA-1B to achieve more robust, class-agnostic segmentation capabilities. SAM excels in producing precise masks across various domains using simple prompts such as points, boxes, and coarse masks. While the boundaries of these masks can closely align with object boundaries, the lack of semantic

*Corresponding Author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-1047



(a) Performance-efficiency comparison of FSS models. The numbers inside the points represent the numbers of parameters.

(b) Comparison with previous generalist and specialist models on various FSS datasets.

Figure 1: Performance comparisons of our approach against previous state-of-the-art methods regarding efficiency and generalized capabilities in Few-shot Semantic Segmentation. Figure 1(a) illustrates our approach's superior performance in efficiency and effectiveness across various model sizes. Figure 1(b) demonstrates the generalizability of our approach across different domains.

understanding and the requirement for manual prompts prevent SAM from being used in automatic semantic segmentation applications.

Recent studies have attempted to automate this process in the Few-shot Semantic Segmentation (FSS), using a few reference images and a fine-grained external backbone network (*e.g.*, DINOv2 [11]) to guide SAM in segmenting target semantic objects. However, these methods face two main challenges: achieving suitable points for precise and full coverage of the target object, and handling the ambiguity of SAM-generated masks, from partial to complete coverage. Specifically, they either utilize the most similar candidate point prompts for iterative mask generation and refinement [12], or build a restrictively selected set of point prompts for heuristically weighted mask merging based on manually designed metrics [13], outperforming both previous specialist methods [14–19] and generalist methods without SAM [20, 21]. However, these methods overlooked the underlying relationships between points (derived from fine-grained features) and masks (generated by SAM in a coarse-grained manner). This oversight led to low efficiency (as indicated in Fig. 1(a)) and limited automation capabilities. Alignment between these two types of granularity could uncover the potential of simple decision-making on masks, which can eliminate redundant refinement and manual hyperparameter selection for complicated metrics.

In this paper, we explicitly explore the relationship between point prompts and corresponding masks from SAM, and present a simple yet effective parameter-free framework with only one-time mask generation to segment anything semantically, in a graph-based few-shot manner. We first introduce a Positive-Negative Alignment (PNA) module to dynamically select point prompts using foreground and background references from reference images. Unlike existing methods, our approach combines different granularity by constructing a directed graph according to mask coverage over points. Then, we perform connectivity analysis on the constructed graph to obtain several weakly connected components as automatic clustering of point prompts, which bridges points and masks as well as fine-grained and coarse-grained features. To mitigate the inevitable introduction of false positives in the PNA module, we further leverage weakly connected component clusters and limited semantic information in selected points, to more accurately filter and merge masks that mismatch in different granularities. In particular, our proposed method involves two post-gating based on weakly connected clusters: the positive gating retains masks capturing a greater proportion of potential target areas, while the overshooting gating screens out outlier points near object boundary, with coverage self-consistency consideration.

Extensive experimental analysis on Few-shot Semantic Segmentation demonstrates both the efficiency and effectiveness of our approach, as shown in Fig. 1(b). We first conduct the experiments on generalized FSS datasets, including Pascal-5^i [22], COCO-20^i [23], FSS-1000 [24] and LVIS-92^i [13]. Our approach surpasses existing state-of-the-art approaches on these datasets, with 5.8% and 2.2%

https://doi.org/10.52202/079017-1047

of improvement respectively on more challenging COCO-20^i and LVIS-92^i. As for the challenging One-shot Part Segmentation, our approach still exceeds previous methods with 1.6% of mIoU on both PACO-Part and PASCAL-Part. Furthermore, to demonstrate the ability of our approach across different domains, we perform an evaluation on several specific datasets, including Deepglobe [25], ISIC [26], and iSAID-5^i [27]. The proposed approach establishes new state-of-the-art performance on mIoU with 49.5% on Deepglobe, 48.7% on ISIC, and 47.3% on iSAID-5^i.

Overall, our contributions are summarized as follows:

- We present, to our knowledge, the first graph-based approach for SAM-based few-shot semantic segmentation, modeling the relationship of SAM-generated masks in an automatic clustering manner.
- We propose a positive-negative alignment module and a post-gating strategy based on the weakly connected graph components, enabling a hyperparameter-free pipeline.
- Extensive experimental comparisons and analysis across several datasets over various settings show the effectiveness and efficiency of the proposed method.

2 Related Work

Few-shot semantic segmentation. Few-shot Semantic Segmentation (FSS) [22] aims to segment the target object using only a limited number of annotated reference samples for guidance. Previous FSS methods are mainly categorized into two types, namely the methods based on prototype matching [28–34] and methods based on pixel-wise matching [35–40]. The methods based on prototype matching, *e.g.* PFENet [31], BAM [41], SSP [42], use the Mask Average Pooling operation from SGOne [43] to generate a prototype as a global representation of the reference features, and compare the target features with the prototypes. The methods based on pixel-wise matching compute the correlation of all pixels between target and reference features. Then different methods address the correlations through distinct mechanisms, such as 4D Convolution (*e.g.*, HSNet [14]) and Transformer (*e.g.*, HDMNet [44], AMFormer [45]). Although these specialist models perform significantly on specific tasks, they are prone to overfitting the training samples and often struggle to adapt to domain shifts.

SAM-based semantic segmentation. Recently, Segment Anything Model (SAM) [10] has shown remarkable zero-shot class-agnostic segmentation capabilities using prompts like points, boxes, and coarse masks. However, the coarse-grained feature representation of SAM limits its effectiveness for fine-grained semantic segmentation tasks. Several approaches have been proposed to extend SAM for semantic segmentation. For example, Semantic-SAM [46] jointly train the model on SA-1B and other semantic aware segmentation datasets to enhance granularity. OV-SAM [47] combines SAM and CLIP [48] for open-vocabulary semantic segmentation. Moreover, some methods introduce SAM into FSS tasks. PerSAM and PerSAM-F [12] leverage SAM for personalized segmentation with one-shot guidance. Matcher [13] uses a SAM-based training-free structure, achieving impressive performance in both FSS and One-shot Part Segmentation. VRP-SAM [49] trains an external Visual Reference Prompt Encoder to automatically generate prompts from reference images using points, scribble, box, or masks. However, previous training-free methods struggled to balance performance and efficiency, often relying on excessive external manual hyperparameters.

3 Preliminaries

Few-shot Semantic Segmentation (FSS) aims to segment target objects in an image with a few annotated reference images. Consider a scenario where each group of samples contains a target image x^t and a reference image set $R = \{x_k^r, y_k^r\}_{k=1}^K$ with the size of $H \times W$, where x_k^r and y_k^r mean the k_{th} reference image and its corresponding mask. Focusing on the 1-shot case, where $K = 1$, it begins with a feature extraction backbone network $f_B(\cdot)$, which encodes both x^t and x^r into semantic features F^t and F^r in $\mathbb{R}^{hw \times c}$, where h and w denote the height and width of the feature maps, and c is the feature dimension. Subsequent few-shot processes utilize these feature maps to generate a predicted segmentation $\tilde{y} \in \mathbb{R}^{H \times W}$ for x^t. This prediction is then compared to the Ground Truth (GT) y^t for evaluation.

The Segment Anything Model (SAM) is a generalized foundation segmentation model adept at generating precise masks based on varied prompts of points, boxes, and coarse masks. Built around

https://doi.org/10.52202/079017-1047

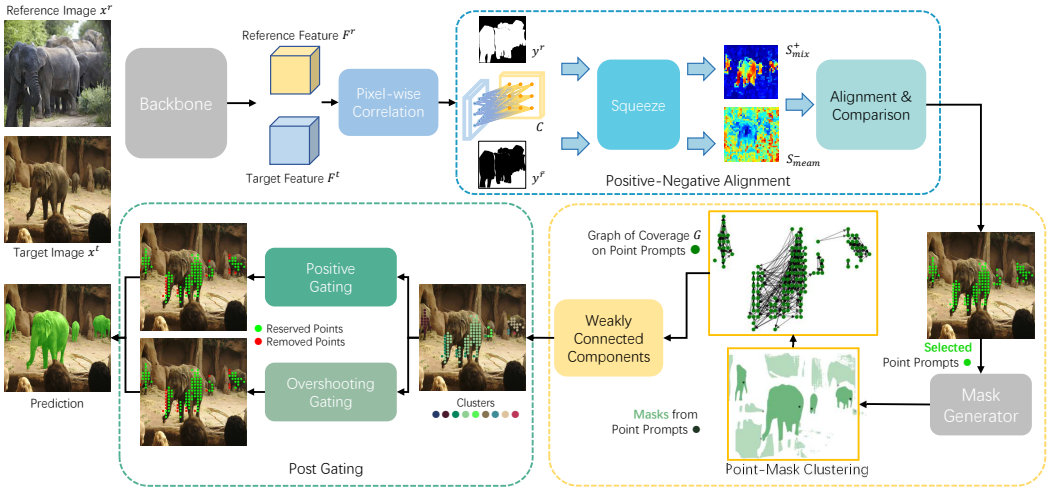


Figure 2: Overview of our approach, where the Positive-Negative Alignment module recognizes the correlation between target features and reference features for point selection, the Point-Mask Clustering module efficiently clusters the points based on the coverage of corresponding masks, and Post-Gating filters out the false-positive masks for generating final prediction.

a core architecture that includes an image encoder, a prompt encoder, and a mask decoder, SAM effectively processes input images x^t and prompts P to produce detailed segmentation masks $\hat{y}$. These masks accurately delineate specific objects or regions within the images, based on the guidance provided by the prompts.

4 Method

Diverging from traditional methods, we use a directed graph to exploit the natural relationships between points and their corresponding masks, representing fine-grained and coarse-grained features, respectively. As shown in Fig. 2, our approach mainly comprises the Positive-Negative Alignment (PNA) module, Point-Mask Clustering (PMC) module, and Post-Gating strategy. The PNA module leverages semantic features from the backbone network to sort pixel-wise correlations into similarity maps, enabling precise point selection. The PMC module then clusters masks based on these selected points, while Post-Gating strategy refines the selection, enhancing the accuracy and reliability of the final prediction.

4.1 Positive-Negative Alignment for Point Selection

The PNA module efficiently selects point prompts to balance the number of points and coverage of target objects. Using the semantic features F^r and F^t from the reference and target images respectively (with *e.g.*, DINOv2 [11]), we get the pixel-wise correlation matrix $C \in \mathbb{R}^{hw \times hw}$:

$$C(i,j) = ReLU\left(\frac{F^t(i) \cdot F^r(j)}{\|F^t(i)\| \cdot \|F^r(j)\|}\right), \tag{1}$$

where $C(i,j)$ represents the similarity between the i-th pixel of target features $F^t(i)$ and the j-th pixel of reference features $F^r(j)$.

To minimize hyperparameter reliance, we leverage background features typically overlooked in FSS, indicated by the negative mask $y^{\tilde{r}} = \neg y^r$ of the reference image. According to y^r and $y^{\tilde{r}}$, we divide C into C^+ and C^- in $\mathbb{R}^{hw \times hw}$ for foreground and background features, respectively. We then introduce two positive similarity maps in mean and max aspects respectively:

$$S^+_{mean}(i) = \frac{\sum_{j=1}^{hw} C^+(i,j)}{\sum_{j=1}^{hw} \mathcal{I}(y^r)_j}, \quad S^+_{max}(i) = max(C^+(i)), \tag{2}$$

where $\mathcal{I}$ resizes y^r to the same resolution as F^r and then flatten it into a vector, $max(\cdot)$ finds the maximum value in the i-th row of C^+. The mean positive similarity map $S^+_{mean} \in \mathbb{R}^{hw}$ captures global similarity towards the reference object but may blur distinct internal features, reducing accuracy for complex objects. In contrast, the max positive similarity map $S^+_{max} \in \mathbb{R}^{hw}$ focuses on the most similar regions, enhancing recall but also increasing noise. To maintain distinctiveness while reducing noise, we introduce the mixture similarity map $S^+_{mix} = S^+_{mean} \odot S^+_{max}$ using the Hadamard product. This method boosts target region distinctiveness by merging the strengths of both maps, while diminishing noise through the more stable global similarity.

To select prompt points, we also use the mean negative similarity map S^-_{mean}, which reflects background similarity, noting that similar objects typically share higher background similarity values. We then align S^+_{mix} and S^-_{mean} by min-max normalization $\mathcal{M}$ to get:

$$S_p(i) = \mathcal{M}(S^+_{mix})(i) \cdot \mathbf{1}_{\{\mathcal{M}(S^+_{mix})(i) > \mathcal{M}(S^-_{mean})(i)\}}, \tag{3}$$

where $S_p \in \mathbb{R}^{hw}$ is the filtered map for point selection, and $\mathbf{1}_{\{\cdot\}}$ is 1 if the condition is true and 0 otherwise. Although we minimize false negatives, noise points remain. To select suitable points from S_p without hyperparameters, we define the sum of elements in S_p as the number N of points to be selected. We then pick the N highest-value points from S_p as the point prompt set $\boldsymbol{P} = \{P_l\}^N_{l=1}$.

4.2 Point-Mask Clustering with Graph Connectivity

We utilize point prompts from $\boldsymbol{P}$ to generate masks with SAM. Each point P_l in $\boldsymbol{P}$ corresponds to a unique mask $\hat{y}_l \in \mathbb{R}^{H \times W}$. As our point selection strategy prioritizes the coverage of objects, false-negative masks are unavoidable. Moreover, mask coverage can vary significantly within the same region, ranging from partial to full object coverage. This necessitates understanding the internal relationships among coarse-grained masks and points from fine-grained feature comparison to ensure those covering the same target are accurately gathered.

To address this, we design the Point-Mask Clustering (PMC) module, which clusters points and their corresponding masks based on mask coverage over points. Following the principles of efficiency and automation, the PMC module is based on a directed graph $G = (V, E)$ with the vertex v_l in V representing point P_l and its corresponding mask $\hat{y}_l$. Edges in E are established based on mask coverage over other points; an edge $e_{l,m}$ exists if mask $\hat{y}_l$ covers points P_m (with $m \neq l$). Specifically, we do not establish edges for masks covering their corresponding points to avoid creating loops.

The graph G is a directed simple graph, allowing us to cluster vertices by identifying weakly connected components. This clustering process is hyperparameter-free, ensuring that every pair of vertices $u, v \in V$ within the same component has a directed path between them. Each weakly connected component encompasses a set of points $\hat{P}_p$ (with $P = \{\hat{P}_p\}$) that are all covered by the union of their masks in $\hat{M}_p$, where p indexes the clusters.

The advanced SAM plays a crucial role in maintaining the precision of the generated masks. The precision of high-quality masks typically ensures non-overlapping between masks and prompting points of adjacent regions, especially those of different categories. This is the precondition for the efficacy of our PMC module, as even slight errors could significantly impact the clustering accuracy.

4.3 Post-Gating on Weakly Connected Components

Our PNA module, while efficient in selecting points, inadvertently includes false positives, as detailed in Sec. 4.1. To mitigate this, we further develop two gating strategies targeting distinct types of false positives based on clusters formed from weakly connected components.

Positive gating. Despite the method in Sec. 4.1 diminishing the noise points outside the target region, there are still a few remaining noise points. These issues may have minimal impact on traditional segmentation methods, but under the SAM framework, masks derived from these noise points can significantly degrade accuracy. Moreover, some clusters of masks may extend beyond their intended target regions due to inaccuracies in SAM-generated masks or because the targeted object is part of a larger entity. Thus, we propose a Positive Gating strategy to address these issues.

This strategy prioritizes mask effectiveness by assessing whether a mask contains more positive than negative pixels, thereby facilitating a specialized designed mask growth for final prediction. The

https://doi.org/10.52202/079017-1047

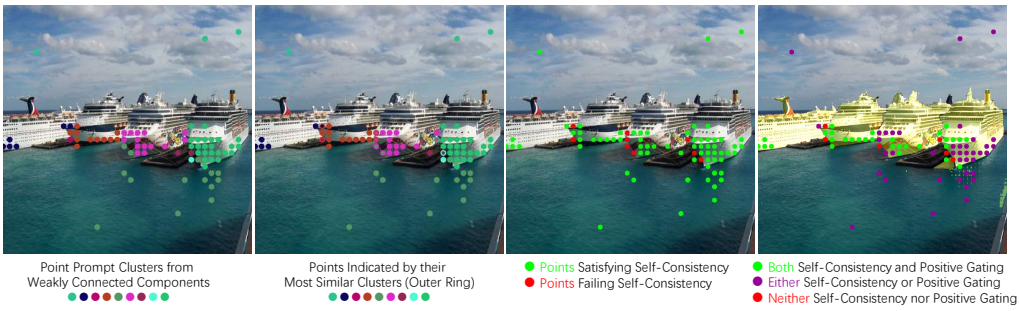


Figure 3: Illustration of the Overshooting Gating strategy. The outer ring of points in the second image indicates the most similar cluster of corresponding points, *i.e.*, points with different outside and inside colors do not satisfy the self-consistency.

focus of mask growth is to enhance coverage of the target area rather than multiple objects, while minimizing background inclusion. Firstly, this method utilizes a parameter-free gating mechanism that discriminates between pixel polarities, based on the positive and negative similarity maps, S^{+}_{mean} and S^{-}_{mean}, as described in Sec. 4.1. To achieve this, we utilize S^{+}_{mean} and S^{-}_{mean}, along with the median of S^{+}_{mean} (*i.e.*, the midpoint between the maximum and minimum values of S^{+}_{mean}), to constructs the polarity map $\bar{R}$ as follows:

$$\bar{R}(i) = \begin{cases} 1, & S^{+}_{mean}(i) \times S^{+}_{mean}(i) > s_{mid} \times S^{-}_{mean}(i), \\ -1, & else. \end{cases} \tag{4}$$

Then, using the polarity map $\bar{R}$, we calculate the number of positive pixels of the l^{th} mask as follows:

$$s^{+}_{l} = \sum_{i=1}^{hw} \bar{R}(i) \odot \mathcal{I}(\hat{y}_l)(i), \tag{5}$$

where $\mathcal{I}$ resizes and flattens $\hat{y}_l$ to the feature map dimensions. Subsequently, for each cluster $\hat{M}_p$ of weakly connected components, we sort the masks according to the ratio of positive pixel numbers to their areas. The indices of these sorted masks are denoted by Q. We then initialize a blank pseudo mask $\ddot{y}_p \in \mathbb{R}^{H \times W}$ and a set of positive points P^{+}. Following this, we apply a Mask Growth algorithm as outlined in Sec. A.1 and Alg. 1 for maintaining positive masks. This algorithm iteratively evaluates whether the region of $\hat{y}_q$ outside the pseudo mask $\ddot{y}_p$ is positive, updates $\ddot{y}_p$ with the identified positive mask, and adds its corresponding point into P^{+}.

Overshooting gating. The fine-grained semantic features from $f(\cdot)$ are reliable for locating target objects, yet the point coverage of the target areas varies, leading to both under-coverage and over-coverage. SAM effectively addresses under-coverage; however, over-coverage, which extends beyond target boundaries, often produces false-positive masks. These overshooting points, while semantically similar to the target areas in F^t, typically derive masks that cover areas outside the target, resulting in a mismatch of representations between the granularity of points and masks. Thus, these points cannot be clustered with points inside the target areas.

Hence, we devise an overshooting gating strategy with consideration of self-consistency to eliminate overshooting points and their associated masks. As shown in Fig. 3, We assess the similarity between the features of each point P_l and the union mask $\hat{y}_p \in \mathbb{R}^{H \times W}$ from each mask cluster $\hat{M}_p$. The similarity computation for estimating **s**elf-**c**onsistency is performed as follows:

$$s^{sc}(l,p) = \frac{\sum_{i=1}^{hw} Sim(F^t(P_l), (F^t \odot \mathcal{I}(\hat{y}_p))(i))}{\sum_{i=1}^{hw} \mathcal{I}(\hat{y}_p) \cdot dist(l,p)}, \tag{6}$$

where $Sim(\cdot,\cdot)$ refers to the correlation calculation mentioned in Eq. 1. We introduce an external function $dist(\cdot,\cdot)$ to measure the distance in F^t between each point P_l and the nearest selected point in $\hat{P}_p$. This measure helps confine comparison to neighboring clusters, minimizing interference from other instances. We then identify the cluster most similar to the points and retain those in the point set P^{sc} that are more similar to their respective clusters.

https://doi.org/10.52202/079017-1047

Table 1: Performance on Few-shot Semantic Segmentation datasets of Pascal-5^i, COCO-20^i, FSS-1000, and LVIS-92^i. Gray means the in-domain trained results. The best results are shown in **bold**.

Methods	Pascal-5^i		COCO-20^i		FSS-1000		LVIS-92^i	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
specialist model								
HSNet [14][CVPR21]	66.2	70.4	41.2	49.5	86.5	88.5	17.4	22.9
VAT [50][ECCV22]	67.9	72.0	41.3	47.9	90.3	90.8	18.5	22.7
HDMNet [44][CVPR23]	69.4	71.8	50.0	56.0	-	-	-	-
AMFormer [45][NeurIPS23]	70.7	73.6	51.0	57.3	-	-	-	-
generalist model								
PerSAM [12][ICLR24]	43.1	-	23.0	-	71.2	-	11.5	-
PerSAM-F [12][ICLR24]	48.5	-	23.5	-	75.6	-	12.3	-
Matcher [13][ICLR24]	68.1	74.0	52.7	60.7	87.0	**89.6**	33.0	40.0
VRP-SAM [49][CVPR24]	71.9	-	53.9	-	-	-	-	-
Ours	**72.1**	**82.6**	**58.7**	**66.8**	**88.0**	88.9	**35.2**	**44.2**

Table 2: Performance on One-shot Part Segmentation datasets and Cross Domain Few-shot Semantic Segmentation datasets. The best results are shown in **bold**.

Methods	One-shot Part Seg.		Cross Domain FSS					
	PASCAL-Part	PACO-Part	Deepglobe		ISIC		iSAID-5^i	
	1-shot	1-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
specialist model								
HSNet [14][CVPR21]	32.4	22.6	29.7	35.1	31.2	35.1	34.1	40.4
DRA [51][CVPR24]	-	-	41.3	50.1	40.8	48.9	-	-
FRINet [52][TGRS23]	-	-	-	-	-	-	42.6	44.5
generalist model								
PerSAM [12][ICLR24]	32.5	22.5	31.4	-	23.9	-	19.2	-
PerSAM-F [12][ICLR24]	32.9	22.7	35.0	-	23.6	-	20.3	-
Matcher [13][ICLR24]	42.9	34.7	48.1	50.9	38.6	35.0	33.3	34.3
Ours	**44.5**	**36.3**	**49.5**	**57.7**	**48.7**	**55.2**	**47.1**	**52.4**

Mask Merging. Finally, we obtain two distinct sets of points, namely P^+ and P^{sc}. We then union the masks corresponding to points that are common to both P^+ and P^{sc}. The merged masks form the final prediction, denoted as $\tilde{y}$.

5 Experimental Results

5.1 Datasets

To illustrate the Few-shot Semantic Segmentation ability and generalization capacity, we conduct three types of sub-tasks, *i.e.* standard Few-shot Semantic Segmentation, One-shot Part Segmentation, and Cross Domain Few-shot Semantic Segmentation. The datasets for these tasks are as follows:

Pascal-5^i, **COCO-20^i**, **FSS-1000**, and **LVIS-92^i** are standard FSS datasets. Pascal-5^i [22] is based on the Pascal VOC 2012 [53] and SDS [54]. The 20 classes are separated into 4 folds of 5 classes. COCO-20^i [23] is an 80-class dataset from MSCOCO [55], which has 4 folds with each fold containing 20 classes. FSS-1000 [24] contains 1000 classes. The training, validation, and testing folds contain 520, 240, and 240 classes, respectively. LVIS-92^i [13] is more challenging for evaluating generalist models, which select 920 classes with more than 2 images and divide these classes into 10 folds.

PASCAL-Part and **PACO-Part** [13] are One-shot Part Segmentation datasets. PASCAL-Part [56, 57] contains 56 different object parts in 4 superclasses. PACO-Part is built based on the PACO dataset [58], which has 456 object part classes. The 303 classes with at least 2 samples in PACO-Part are divided into four folds following Matcher [13].

Deepglobe, **ISIC2018**, and **iSAID-5^i** are Cross Domain FSS datasets. The Deepglobe [25] contains satellite images of geographic categories including urban, agriculture, rangeland, forest, water, and

https://doi.org/10.52202/079017-1047

Table 3: Ablation study of Point Selection.

S^+_{mean}	S^+_{max}	S^-_{mean}	Top N	mIoU
✓		✓	✓	53.1
	✓	✓	✓	54.1
✓	✓	✓		56.4
✓	✓		✓	51.5
✓	✓	✓	✓	**58.7**

Table 4: Ablation study of PMC and Post-Gating.

PG Strong	PG Weak	OG Strong	OG Weak	COCO-20^i	LVIS-92^i
				44.0	24.2
	✓			57.1	34.3
✓				57.1	33.9
	✓	✓		56.7	**35.2**
	✓		✓	**58.7**	**35.2**
k-means++				57.5	34.0

Table 5: Ablation study of positive gating on each cluster. M.G. represents the Mask Growth algorithm.

Strategies	M.G.	COCO-20^i	PASCAL-Part
Sum		55.3	39.1
	✓	58.6	44.3
Num		57.1	42.2
	✓	**58.7**	**44.5**

Table 6: Ablation on the strategies of Self-Consistency measurement.

Strategies	mIoU	Δ
None	57.1	0.0
Point Sim.	56.7	-0.4
MAP Sim.	57.7	+0.6
Mean Sim. W/o dist	49.1	-8.0
Mean Sim. (Ours)	**58.7**	**+1.6**

barren. The ISIC2018 [26] is a skin lesion analysis dataset with three classes. The iSAID-5^i [27] evenly split 3 folds for 15 classes based on the remote sensing dataset iSAID [59].

5.2 Implementation Details

Following the settings of PerSAM [12] and Matcher [13] for a fair comparison, we use DINOv2 [11] with a ViT-L/14 [60] as our feature extraction backbone, and SAM [10] with ViT-H as the mask generator. The input image sizes are set to 518×518 for DINOv2 and 1024×1024 for SAM following Matcher [13]. Except for the default hyperparameters of SAM and DINOv2, our approach **does not have any external hyperparameter**. We apply the mean Intersection over Union (mIoU) metric for evaluating the performance. All experiments are conducted on a single NVIDIA RTX2080Ti.

5.3 Comparison with State-Of-The-Arts

Comparison on the standard FSS datasets. We compared our approach with other state-of-the-art specialist and generalist models. As shown in Tab. 1, our approach achieves 72.1% mIoU on the Pascal-5^i dataset and 58.7% mIoU on COCO-20^i dataset, which surpasses all previous specialist and generalist state-of-the-art models. Our approach reaches 35.2% mIoU in the more challenging dataset of LVIS-92^i, with 2.2% of improvement compared to the previous training-free method Matcher. The performance remains competitive on the FSS-1000 compared with specialist models. The 5-shot result of Pascal-5^i, COCO-20^i, and LVIS-92^i further extends the lead, which shows that our approach can effectively handle the few-shot scenario.

Comparison on the One-shot Part Segmentation datasets. The One-shot Part Segmentation tasks evaluate the ability to fetch the target part from the whole object. The results in Tab. 2 show that our approach achieves the mIoU of 44.5% and 36.3% on both datasets of PASCAL-Part and PACO-Part, respectively. Our approach outperforms the state-of-the-art generalist model Matcher with 1.6% on both datasets. Given that Matcher employs **specific hyperparameters** to enhance part segmentation, our superior performance demonstrates the adaptability of our approach across both object and part segmentation contexts.

Comparison on the Cross Domain FSS datasets. The Cross Domain FSS tasks validate the performance on different domains. Our approach achieves state-of-the-art performance in datasets of Deepglobe, ISIC, and iSAID-5^i among other specialist domain models and generalist models. Especially within the context of the skin lesion analysis dataset ISIC and remote sensing dataset iSAID-5^i, our approach outperforms Matcher by margins of 10.1% and 13.8% respectively.

https://doi.org/10.52202/079017-1047

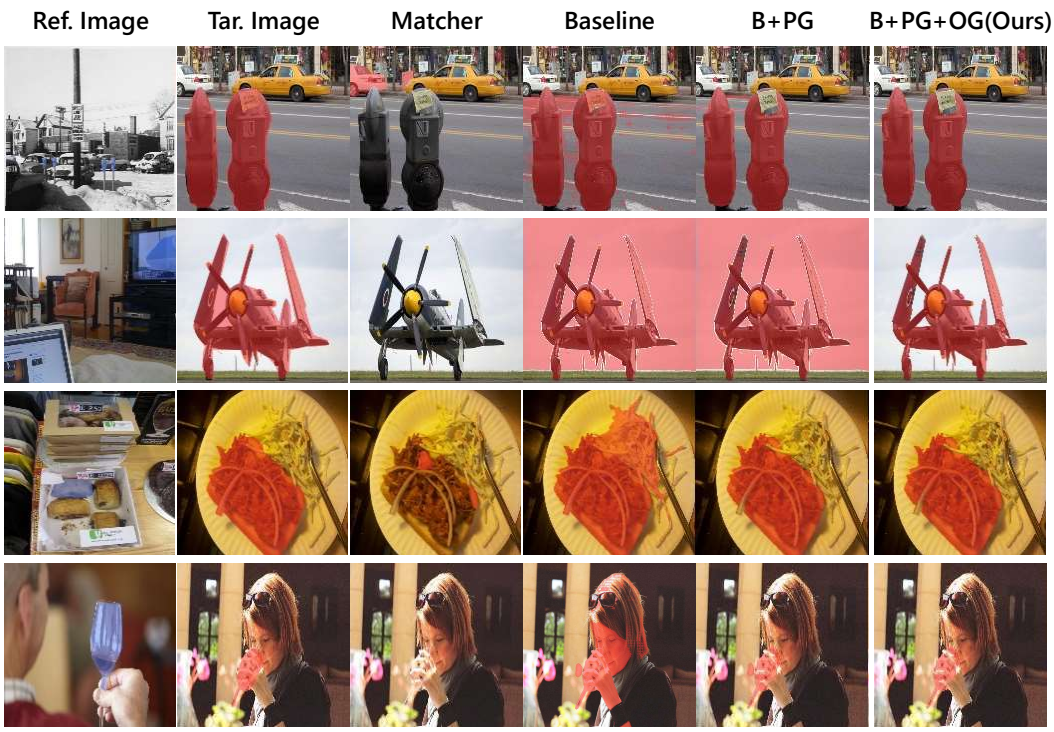


Figure 4: Qualitative analysis of Matcher, Baseline, B+PG, B+PG+OG. B, PG, and OG respectively represent Baseline, Positive Gating, and Overshooting Gating. Masks in ref. image are shown in blue.

5.4 Ablation Study

Point selection. We evaluate the impact of various similarity maps and the parameter-free selection of top N points on performance, as detailed in Sec. 4.1. As shown in Tab. 3, using either S^+_{mean} or S^+_{max} alone leads to a performance drop of up to 5.6% compared to using both. This decline is due to the inherent limitations of S^+_{mean} and S^+_{max} discussed in Sec 4.1. Additionally, the evaluation confirms that picking the top-N points based on similarity, which is parameter-free and requires no additional settings, simplifies the process and increases accuracy by 2.1%.

Clustering method. We compare our PMC module using weakly connected components with the PMC module using strong connected components, which provides finer clustering results. According to our experiment results in Tab. 4, the clusters from weakly connected components provide better performance on COCO-20^i and LVIS-92^i for both gating, as these clusters of masks have ideal coverage of the objects. Simply filtering the masks without clustering-based gating can only achieve 44.0% mIoU on COCO-20^i and 24.2% mIoU on LVIS-92^i, which is significantly lower than the performance achieved with clustering-based gating. Furthermore, our dynamic hyperparameter-free clustering method outperforms the k-means++ with 1.2% on both datasets. Note that k of k-means++ is set to 10 following Matcher [13].

Positive gating. Our approach compares the number of positive points and negative points in $\hat{S}^+$ (Num) to judge whether the mask is positive. We conduct experiments for the strategy of comparing the sum of positive and negative values (Sum). The results in Tab. 5 demonstrate the Num strategy yields better performance, as comparing the number of pixels mitigates the influence of a few excessively high similarity values. Furthermore, the utilization of the Mask Growth algorithm improves both FSS and Part Segmentation performance by carefully retaining the positive regions. However, it weakens the improvement of Num due to their similar effects.

Overshooting gating. Our Overshooting gating aims to filter out the overshooting points closely neighboring to target regions, thus having a less remarkable improvement of 1.6% compared to Positive Gating, as shown in Tab. 6. This performance still surpasses the mean similarity of com-

https://doi.org/10.52202/079017-1047

paring points with regions of clustered points (Point Sim.) or the prototypes from Masked Average Pooling [43] with union masks (MAP Sim.). More importantly, the distance function avoids the gating from 9.6% of performance decline. It ensures each cluster only affects its neighboring points.

5.5 Qualitative Analysis

Here we present the qualitative results of Matcher, Baseline (1st row in Tab. 4), Baseline+PG (2nd row in Tab. 4) and our approach in Fig. 4. The bipartite matching of Matcher has a negative influence when the areas of the target object in reference and target images have significant differences, as shown in the 1st and 3rd rows. The positive gating with clustering filters out the noise masks in the 3rd row, while the Overshooting Gating further removes the masks belonging to overshooting points in the 2nd and 4th rows. More qualitative analyses please refer to the appendix.

6 Conclusion

In this paper, we proposed an efficient, training-free SAM-based FSS approach that requires no external hyperparameters. As an automatic SAM-based semantic segmentation pipeline, our approach balanced candidate points and object coverage in the Positive-Negative Alignment (PNA) module, then used SAM-generated masks in the Point-Mask Clustering (PMC) module to enhance Post Gating. Extensive experiments validated the superior performance of our approach, advancing semantic segmentation without extensive parameter tuning or training.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China under No. 62472033, No. U23A20314, and No. 61972036. J. Jiao is supported by the Royal Society Short Industry Fellowship (SIF\R1\231009) and the Amazon Research Award.

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Recognition and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[5] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12077–12090, 2021.

[6] Z. Zhang, G. Gao, Z. Fang, J. Jiao, and Y. Wei, "Mining unseen classes via regional objectness: A simple baseline for incremental segmentation," *Advances in neural information processing systems*, vol. 35, pp. 24340–24353, 2022.

[7] Z. Zhang, G. Gao, J. Jiao, C. H. Liu, and Y. Wei, "Coinseg: Contrast inter-and intra-class representations for incremental segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 843–853, 2023.

[8] A. Zhang and G. Gao, "Background adaptation with residual modeling for exemplar-free class-incremental semantic segmentation," *Proc. European Conference on Computer Vision*, 2024.

[9] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in neural information processing systems*, vol. 34, pp. 17864–17875, 2021.

[10] A. Kirillov, E. Mintun, and et al., "Segment anything," in *Proc. IEEE International Conference on Computer Vision*, pp. 4015–4026, 2023.

[11] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[12] R. Zhang, Z. Jiang, Z. Guo, S. Yan, J. Pan, H. Dong, Y. Qiao, P. Gao, and H. Li, "Personalize segment anything model with one shot," in *Proc. International Conference on Learning Representations*, 2024.

[13] Y. Liu, M. Zhu, H. Li, H. Chen, X. Wang, and C. Shen, "Matcher: Segment anything with one shot using all-purpose feature matching," in *Proc. International Conference on Learning Representations*, 2024.

[14] J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmentation," in *Proc. IEEE International Conference on Computer Vision*, pp. 6941–6952, 2021.

[15] Q. Xu, W. Zhao, G. Lin, and C. Long, "Self-calibrated cross attention network for few-shot segmentation," in *Proc. IEEE International Conference on Computer Vision*, pp. 655–665, 2023.

[16] Y. Sun, Q. Chen, X. He, J. Wang, H. Feng, J. Han, E. Ding, J. Cheng, Z. Li, and J. Wang, "Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning," *Advances in neural information processing systems*, vol. 35, pp. 37484–37496, 2022.

[17] Y. Liu, N. Liu, Q. Cao, X. Yao, J. Han, and L. Shao, "Learning non-target knowledge for few-shot semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11573–11582, 2022.

[18] Z. Fang, G. Gao, Z. Zhang, and A. Zhang, "Hierarchical context-agnostic network with contrastive feature diversity for one-shot semantic segmentation," *Journal of Visual Communication and Image Representation*, vol. 90, p. 103754, 2023.

[19] G. Gao, Z. Fang, C. Han, Y. Wei, C. H. Liu, and S. Yan, "Drnet: Double recalibration network for few-shot semantic segmentation," *IEEE Transactions on Image Processing*, vol. 31, pp. 6733–6746, 2022.

[20] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang, "Seggpt: Towards segmenting everything in context," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1130–1140, 2023.

[21] X. Wang, W. Wang, Y. Cao, C. Shen, and T. Huang, "Images speak in images: A generalist painter for in-context visual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6830–6839, 2023.

[22] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," in *Proc. British Machine Vision Conference*, pp. 167.1–167.13, 2017.

[23] K. Nguyen and S. Todorovic, "Feature weighting and boosting for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 622–631, 2019.

[24] X. Li, T. Wei, Y. P. Chen, Y.-W. Tai, and C.-K. Tang, "Fss-1000: A 1000-class dataset for few-shot segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2869–2878, 2020.

[25] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 172–181, 2018.

 https://doi.org/10.52202/079017-1047

[26] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1902.03368*, 2019.

[27] X. Yao, Q. Cao, X. Feng, G. Cheng, and J. Han, "Scale-aware detailed matching for few-shot aerial image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[28] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5217–5226, 2019.

[29] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE International Conference on Computer Vision*, pp. 9197–9206, 2019.

[30] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, and J. Kim, "Adaptive prototype learning and allocation for few-shot segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8334–8343, 2021.

[31] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior guided feature enrichment network for few-shot segmentation," *IEEE Transactions on Pattern Recognition and Machine Intelligence*, vol. 44, no. 2, pp. 1050–1065, 2020.

[32] M. Boudiaf, H. Kervadec, Z. I. Masud, P. Piantanida, I. Ben Ayed, and J. Dolz, "Few-shot segmentation without meta-learning: A good transductive inference is all you need?," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13979–13988, 2021.

[33] A. Okazawa, "Interclass prototype relation for few-shot segmentation," in *Proc. European Conference on Computer Vision*, pp. 362–378, 2022.

[34] J.-W. Zhang, Y. Sun, Y. Yang, and W. Chen, "Feature-proxy transformer for few-shot segmentation," in *Advances in neural information processing systems*, pp. 6575–6588, 2022.

[35] X. Shi, D. Wei, Y. Zhang, D. Lu, M. Ning, J. Chen, K. Ma, and Y. Zheng, "Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation," in *Proc. European Conference on Computer Vision*, pp. 151–168, 2022.

[36] H. Liu, P. Peng, T. Chen, Q. Wang, Y. Yao, and X.-S. Hua, "Fecanet: Boosting few-shot semantic segmentation with feature-enhanced context-aware network," vol. 25, pp. 8580–8592, 2023.

[37] Y. Wang, R. Sun, and T. Zhang, "Rethinking the correlation in few-shot segmentation: A buoys view," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7183–7192, 2023.

[38] Y. Liu, N. Liu, X. Yao, and J. Han, "Intermediate prototype mining transformer for few-shot semantic segmentation," vol. 35, pp. 38020–38031, 2022.

[39] S. Jiao, G. Zhang, S. Navasardyan, L. Chen, Y. Zhao, Y. Wei, and H. Shi, "Mask matching transformer for few-shot segmentation," *Advances in neural information processing systems*, vol. 35, pp. 823–836, 2022.

[40] G. Zhang, G. Kang, Y. Yang, and Y. Wei, "Few-shot segmentation via cycle-consistent transformer," *Advances in neural information processing systems*, vol. 34, pp. 21984–21996, 2021.

[41] C. Lang, G. Cheng, B. Tu, and J. Han, "Learning what not to segment: A new perspective on few-shot segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8057–8067, 2022.

[42] Q. Fan, W. Pei, Y.-W. Tai, and C.-K. Tang, "Self-support few-shot semantic segmentation," in *Proc. European Conference on Computer Vision*, pp. 701–719, 2022.

[43] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "Sg-one: Similarity guidance network for one-shot semantic segmentation," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3855–3865, 2020.

[44] B. Peng, Z. Tian, X. Wu, C. Wang, S. Liu, J. Su, and J. Jia, "Hierarchical dense correlation distillation for few-shot segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 23641–23651, 2023.

[45] Y. Wang, N. Luo, and T. Zhang, "Focus on query: Adversarial mining transformer for few-shot segmentation," *Advances in neural information processing systems*, vol. 36, pp. 31524–31542, 2023.

[46] F. Li, H. Zhang, P. Sun, X. Zou, S. Liu, J. Yang, C. Li, L. Zhang, and J. Gao, "Semantic-sam: Segment and recognize anything at any granularity," *arXiv preprint arXiv:2307.04767*, 2023.

[47] H. Yuan, X. Li, C. Zhou, Y. Li, K. Chen, and C. C. Loy, "Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively," *arXiv preprint*, 2024.

[48] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[49] Y. Sun, J. Chen, S. Zhang, X. Zhang, Q. Chen, G. Zhang, E. Ding, J. Wang, and Z. Li, "Vrp-sam: Sam with visual reference prompt," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[50] S. Hong, S. Cho, J. Nam, S. Lin, and S. Kim, "Cost aggregation with 4d convolutional swin transformer for few-shot segmentation," in *Proc. European Conference on Computer Vision*, pp. 108–126, 2022.

[51] J. Su, Q. Fan, G. Lu, F. Chen, and W. Pei, "Domain-rectifying adapter for cross-domain few-shot segmentation," 2024.

[52] Q. Cao, Y. Chen, C. Ma, and X. Yang, "Few-shot rotation-invariant aerial image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2023.

[53] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal on Computer Vision*, vol. 88, pp. 303–338, 2010.

[54] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. IEEE International Conference on Computer Vision*, pp. 991–998, 2011.

[55] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. European Conference on Computer Vision*, pp. 740–755, 2014.

[56] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1971–1978, 2014.

[57] K. Morabia, J. Arora, and T. Vijaykumar, "Attention-based joint detection of object and semantic part," *arXiv preprint arXiv:2007.02419*, 2020.

[58] V. Ramanathan, A. Kalia, V. Petrovic, Y. Wen, B. Zheng, B. Guo, R. Wang, A. Marquez, R. Kovvuri, A. Kadian, *et al.*, "Paco: Parts and attributes of common objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7141–7151, 2023.

[59] S. Waqas Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. Shahbaz Khan, F. Zhu, L. Shao, G.-S. Xia, and X. Bai, "isaid: A large-scale dataset for instance segmentation in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 28–37, 2019.

[60] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *Proc. International Conference on Learning Representations*, 2020.

 https://doi.org/10.52202/079017-1047

A Appendix / supplemental material

A.1 More Details for Mask Growth Algorithm

We mention the Mask Growth algorithm in Sec. 4.3. The Mask Growth algorithm is designed for each cluster of masks $\hat{M}_{weak,p}$. The details of the algorithm are shown in Alg. 1. We first initialize an empty set P^+ and a blank pseudo mask $\ddot{y}_p$. Then, we start an iterative process and get the current mask $\hat{y}_q$ based on the sorted sequence of indices Q. The parts of the current mask $\hat{y}_q$ overlapping with $\ddot{y}_p$ are removed. We compute the positive value s_q^+ of the remaining parts. If s_q^+ is positive, the mask $\hat{y}_q$ is updated into the $\ddot{y}_p$ and its corresponding point P_q is added into P^+. As soon as the iterative process finishes, the set of positive points P^+ is established.

Algorithm 1 Mask Growth for each cluster

Input: $\hat{M}_p, \ddot{y}_p, Q, P^+$
  for $n = 1$ to $|Q|$ **do**
    $q \leftarrow Q(n)$
    $\hat{y}_q \leftarrow \hat{M}_p(q)$
    $\hat{y}_q = \hat{y}_q \& \sim \ddot{y}_p$
    $s_q^+ \leftarrow \sum_{i=1}^{hw} \hat{S}^+(i) \odot \mathcal{I}(\hat{y}_q)(i)$
    if $s_q^+ > 0$ **then**
      Add P_q to P^+.
      $\ddot{y}_p = \hat{y}_q \vee \ddot{y}_p$
    end if
  end for
Output: P^+

A.2 Limitations

Our approach has impressive performance on Few-shot Semantic Segmentation tasks. However, due to the resolution of features F^t from DINOv2 not aligning with the required resolution for prompting the SAM, we directly map the coordinates of points in F^t to coordinates for prompting. This results in coordinate bias for small objects, as the gap between neighboring points can reach approximately 28 pixels. Our future work will focus on locating small objects.

A.3 Societal Impacts

As a completely automatic SAM-based few-shot semantic segmentation approach without external hyperparameters, our method is capable of handling various scenarios of semantic segmentation, as demonstrated by our extensive experiments. The efficiency and generalizability of our method ensure a wide range of applications. Furthermore, since our training-free method is constructed upon the widely used open-source foundation models, we have not identified the negative societal impact to date.

A.4 Details of Current SAM-based FSS Methods

Our approach aims to address several issues present in previous SAM-based FSS methods to achieve an automatic SAM-based model. These issues include the requirement of excessive external hyperparameters, overusing the mask generator of SAM, prolonged inference times, etc. The Tab. 7 shows the difference between our approach and previous SAM-based FSS methods. Fig. 5 shows the difference in using SAM as the mask generator between our approach and previous methods. Fig. 5(a) presents the iterative refinement of PerSAM, which involves generating masks from SAM 3 times. Fig. 5(b) exhibits that Matcher introduces an external Automatic Mask Generator, which automatically prompts for generating all mask proposals in the image. Our approach in Fig. 5(c) only utilizes the standard Mask Generator of SAM and generates the masks with our prompts only once.

Table 7: Details of the current SAM-based FSS methods.

Methods	PerSAM	PerSAM-F	Matcher	VRP-SAM	Ours
Training-free	✓		✓		✓
External-hyperparameters-free	✓				✓
Once mask generation				✓	✓
Inference speed (s/img)	1.43	16.5	12.7	N/A	1.88
COCO-20^i mIoU	23.0	23.5	52.7	53.9	58.7

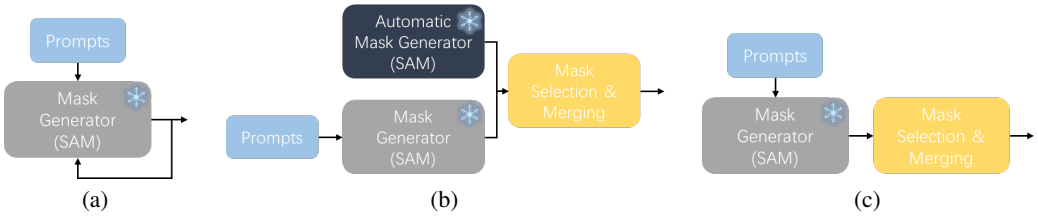


Figure 5: Comparison of the pipeline between the previous methods and our approach. (a) PerSAM [12] iteratively uses the Mask Generator to refine the mask. (b) Matcher [13] introduced an external Automatic Mask Generator [10] with automatic prompting to excessively generate masks from the whole image. (c) The effectiveness of the PMC module and Post-Gating ensures that our approach uses Mask Generator with our prompts only once.

A.5 Discussion of SAM

A.5.1 Features from ViT Encoder of SAM

Previous state-of-the-art generalist FSS methods [] use DINOv2 or ResNet-50, instead of the default ViT encoder of SAM, for fine-grained features. We visualize the representative samples of Pascal-5^i in Fig. 6. The 3rd column of maps represents the self-similarity of the F^t_{SAM}. We introduce the 3×3 average pooling for F^t_{SAM} followed by computing the cosine similarity between the pooled features and F^t_{SAM}. The maps illustrate that F^t_{SAM} can accurately identify the regions of objects within the image, where the features within each object region are nearly identical, while features between neighboring different objects are distinct.

Although the characteristics of F^t_{SAM} ensure the generation of high-quality masks, the coarse-grained features are not suitable for locating the objects, as shown in the 4th column. The similarity between F^t_{SAM} and F^r_{SAM} cannot effectively distinguish the target object well compared to S^+_{mean} from DINOv2. Therefore, we follow the previous methods using DINOv2 for fine-grained features.

A.5.2 Masks analysis for Point-Mask Clustering

Our Point-Mask Clustering module introduces a parameter-free clustering method by constructing a graph of coverage. The effectiveness of the method primarily relies on the high-quality masks, whose boundaries mostly align with the object boundaries. We roughly analyze the coverage of masks generated from the points in the ground truth foreground region using 4000 samples from Pascal-5^i. In particular, we get the union of masks from the foreground points as $\hat{y}_{fore}$, and visualize three distributions, including the distribution of IoU between $\hat{y}_{fore}$ and union of masks from background points $\hat{y}_{back}$ in Fig. 7, the ratio between the number of background points and all points covered by the $\hat{y}_{fore}$ in Fig. 8, the number of background points covered by $\hat{y}_{fore}$ in Fig. 9. The distribution charts demonstrate that most of the samples have acceptable coverage on the error points for Point-Mask Clustering. Given the limited precision of ground truth annotations, the analysis is for reference only. The effectiveness of our Point-Mask Clustering is validated in our ablation study in Tab. 4.

https://doi.org/10.52202/079017-1047

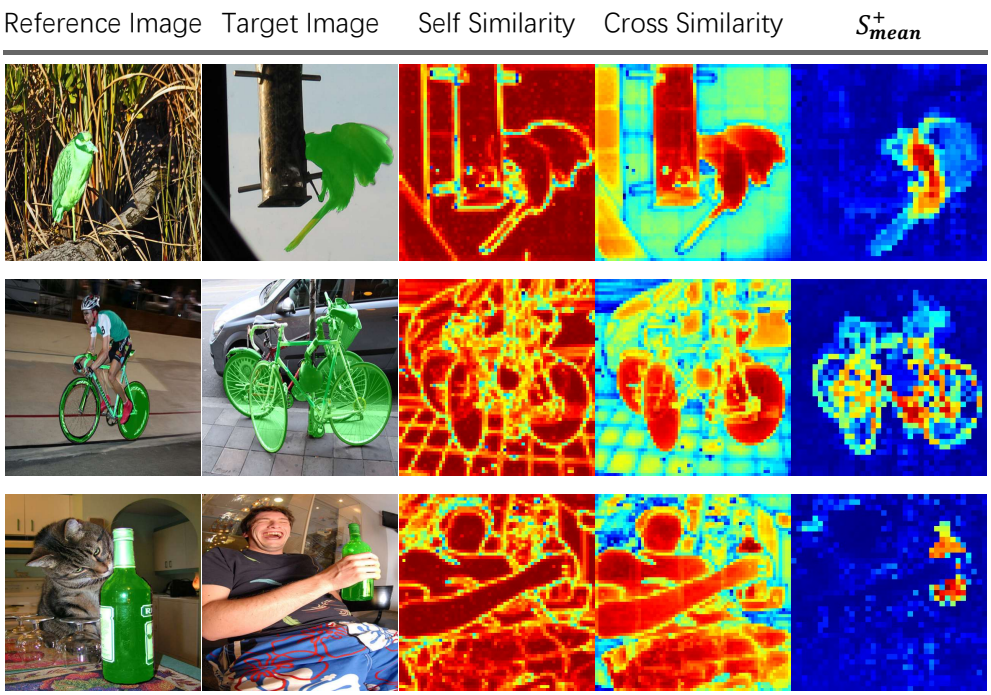


Figure 6: Analysis of the features from default ViT encoder of SAM.

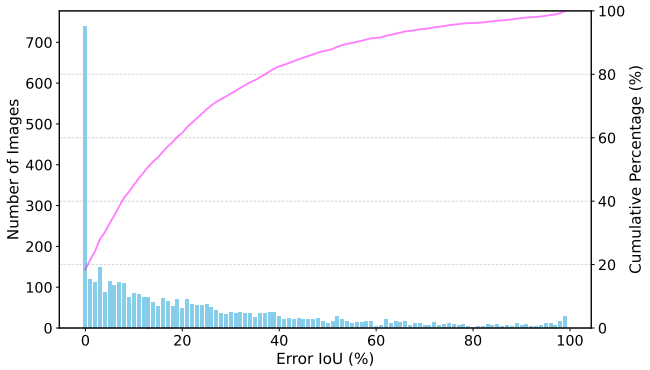


Figure 7: The distribution of IoU between masks from foreground points and from background points.

A.6 Additional Experiment Results.

A.6.1 Performance of Different Foundation Model Sizes

Tab. 8 shows the experiment results of our approach with different scales of SAM and DINOv2. Compared to the previous training-free method Matcher, our approach still achieves a better performance with SAM-Large and DINOv2-Base. The fair comparison with SAM-Huge and DINOv2-Large further demonstrates the effectiveness of our approach.

A.6.2 Detailed Results of Evaluation Datasets

We present the detailed results on different Few-shot Semantic Segmentation datasets, including Pascal-5^i in Tab. 9, COCO-20^i in Tab. 10, LVIS-92^i in Tab. 11, PASCAL-Part and PACO-Part in

https://doi.org/10.52202/079017-1047

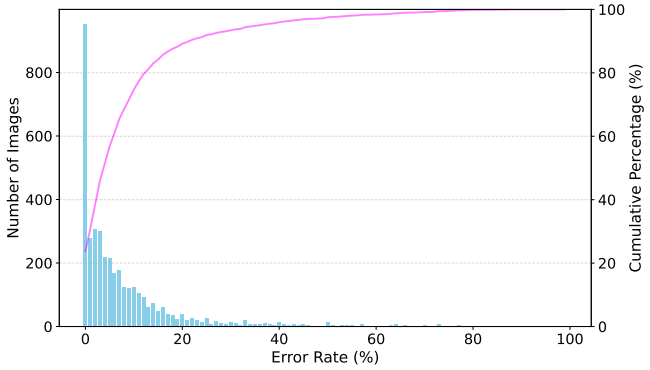


Figure 8: The distribution of the ratio between the number of background points and all points covered by the masks from foreground points.

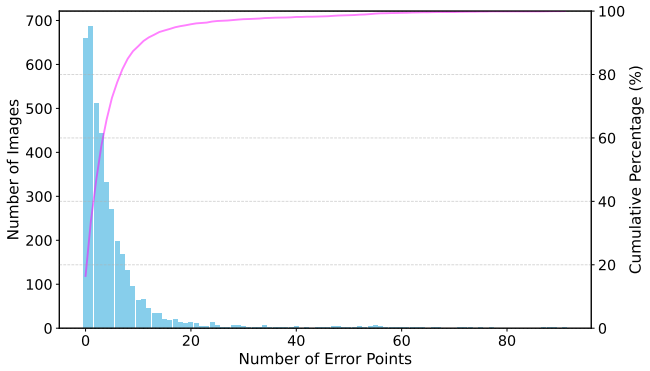


Figure 9: The distribution of the number of background points covered by the masks from foreground points.

Tab. 12, iSAID-5^i in Tab. 13. The results show that our approach has remarkable performance in each fold of the datasets, demonstrating its generalized effectiveness in various scenarios.

A.6.3 Multiple Random Seeds Experiment

Previous state-of-the-art methods, including both specialist methods and generalist methods, typically do not conduct multiple random seed experiments to evaluate the robustness. To demonstrate the robustness of our approach, we randomly set 5 different random seeds and conducted the experiments on the datasets that were not fully evaluated in the standard evaluation. As shown in Fig. 10, despite variations in random seeds, our approach consistently exhibits better performance compared to previous methods that were not evaluated with random seeds.

A.7 Additional Ablation Study

A.7.1 Ablation Study of Pivots for Positive Gating

We apply both s_{mid} and S^{-}_{mean} as the pivots for Positive Gating in Sec. 4.3, aiming to leverage both the pivots from the S^{+}_{mean} itself and the negative similarity. As shown in Tab. 14, combining these two pivots for Positive Gating yields a significant improvement compared to using only one pivot. Moreover, the combination method of $\times$ shows a 0.3% mIoU enhancement compared to $+$.

https://doi.org/10.52202/079017-1047

Table 8: Evaluation of our approach with different sizes of SAM and DINOv2.

Methods	SAM	DINOv2	Params.	COCO-20^i	FSS-1000	LVIS-92^i
Matcher	huge	large	945M	52.7	87.0	31.4
Ours	base	base	180M	53.7	85.6	31.1
	large	base	399M	55.5	87.5	31.7
	large	large	617M	58.0	87.8	35.1
	huge	large	945M	58.7	88.0	35.2

Table 9: Detail results of Pascal-5^i.

Methods	Pascal-5^i 1-shot					Pascal-5^i 5-shot				
	fold0	fold1	fold2	fold3	mean	fold0	fold1	fold2	fold3	mean
AMFormer	71.3	76.7	70.7	63.9	70.7	74.4	78.5	74.3	67.2	73.6
Matcher	67.7	70.7	66.9	67.0	68.1	71.4	77.5	74.1	72.8	74.0
Ours	71.1	75.7	69.2	73.3	72.1	81.5	86.3	79.7	82.9	82.6

Table 10: Detail results of COCO-20^i.

Methods	COCO-20^i 1-shot					COCO-20^i 5-shot				
	fold0	fold1	fold2	fold3	mean	fold0	fold1	fold2	fold3	mean
AMFormer	44.9	55.8	52.7	50.6	51.0	52.0	61.9	57.4	57.9	57.3
Matcher	52.7	53.5	52.6	52.1	52.7	60.1	62.7	60.9	59.2	60.7
Ours	56.6	61.4	59.6	57.1	58.7	67.1	69.4	66.0	64.8	66.8

Table 11: Detail results of LVIS-92^i.

Methods	LVIS-92^i 1-shot										
	fold0	fold1	fold2	fold3	fold4	fold5	fold6	fold7	fold8	fold9	mean
Matcher	31.4	30.9	33.7	38.1	30.5	32.5	35.9	34.2	33.0	29.7	31.4
Ours	30.9	37.9	37.1	39.6	31.2	36.4	39.1	35.7	32.3	31.5	35.2
Methods	LVIS-92^i 5-shot										
	fold0	fold1	fold2	fold3	fold4	fold5	fold6	fold7	fold8	fold9	mean
Matcher	37.0	36.6	47.3	39.1	37.1	41.8	42.7	37.7	37.9	43.3	40.0
Ours	42.1	38.4	50.0	42.5	42.0	46.5	46.4	41.5	43.7	48.4	44.2

Table 12: Detail results of PASCAL-Part and PACO-Part.

Methods	PASCAL-Part					PACO-Part				
	animals	indoor	person	vehicles	mean	fold0	fold1	fold2	fold3	mean
HSNet	21.2	53.0	20.2	35.1	32.4	20.8	21.3	25.5	22.6	22.6
Matcher	37.1	56.3	32.4	45.7	42.9	32.7	35.6	36.5	34.1	34.7
Ours	33.2	59.6	35.2	50.1	44.5	33.4	34.9	39.7	37.0	36.3

Table 13: Detail results of iSAID-5^i.

Methods	iSAID-5^i 1-shot				iSAID-5^i 5-shot			
	fold0	fold1	fold2	mean	fold0	fold1	fold2	mean
FRINet	46.5	36.9	43.9	42.6	48.9	38.1	46.5	44.5
Matcher	37.3	23.8	38.8	33.3	38.3	24.0	40.6	34.3
Ours	53.4	36.8	51.2	47.1	59.3	39.9	58.0	52.4

https://doi.org/10.52202/079017-1047

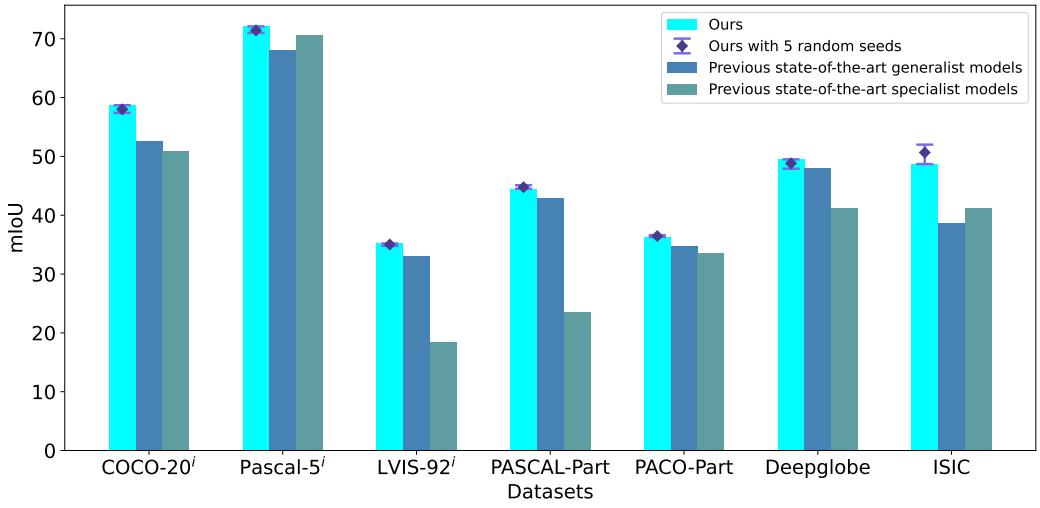


Figure 10: Results of our approach in multiple random seeds experiment. The bars in the chart represent the result under the previous standard evaluation. The error bar depicts the boundaries of our performance.

Table 14: Ablation study of pivots and combination operations in Positive Gating.

Pivots s_{mid}	Pivots S^-_{mean}	Combination	COCO-20^i
✓			44.5
	✓		51.0
✓	✓	+	56.8
✓	✓	×	57.1

Table 15: Ablation study of different strategies for Positive Gating of the masks.

Strategies	COCO-20^i	LVIS-92^i	PASCAL-Part	PACO-Part
Union	59.4	36.1	40.1	31.6
Mask Growth	58.7	35.2	44.5	36.3

A.7.2 Ablation Study of Other Strategies for Positive Gating

In Sec. 4.3 and Sec. A.1, we introduce the Mask Growth algorithm as our strategy for judging whether the mask is positive. We compare the strategy to separately judging each mask in Tab. 5 and judging the union mask of each cluster in Tab. 15. While simply judging the union mask shows better performance on COCO-20^i and LVIS-92^i that require complete coverage of objects, its performance on One-shot Part Segmentation has a significant decline. Considering the generalizability of our approach, we select the Mask Growth algorithm as our strategy.

A.8 Additional Qualitative Analysis

We conduct additional qualitative analysis to better present the result of our approach. Fig. 11 further compare the Matcher, Baseline, B+PJ, and B+PJ+OJ (Ours) following Sec. 5.5. Fig. 12 illustrate the intermediate contents in the Post Gating. Moreover, we provide additional visualization results of standard FSS in Fig. 13, One-shot Part Segmentation in Fig. 14, and Cross Domain FSS in Fig. 15. These qualitative results demonstrate the effectiveness of our approach. Notably, some of the results are even better than the corresponding annotations.

https://doi.org/10.52202/079017-1047

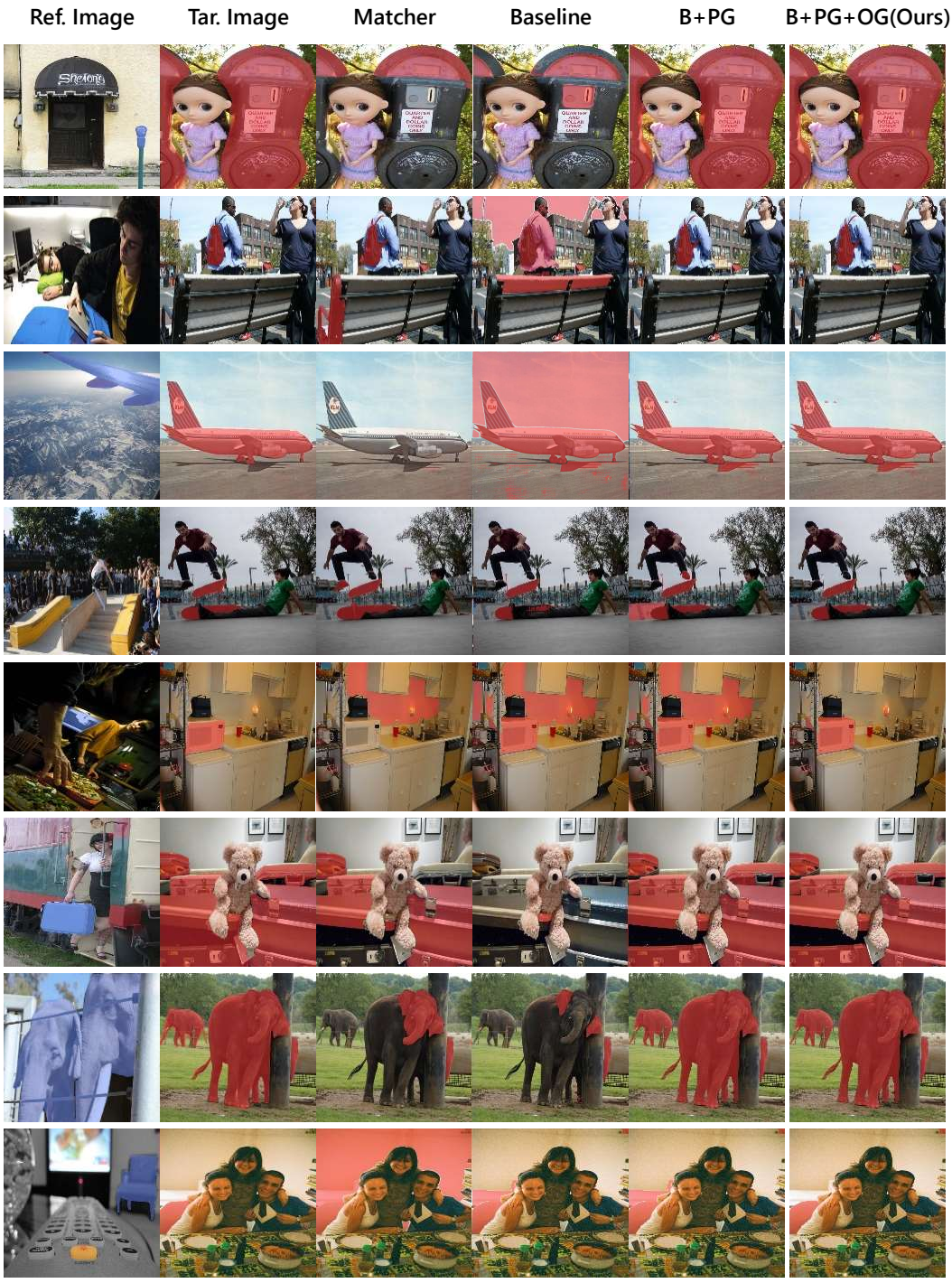


Figure 11: More Qualitative results on COCO-20^i for comparison among Matcher, Baseline, B+PG, B+PG+OG.

https://doi.org/10.52202/079017-1047

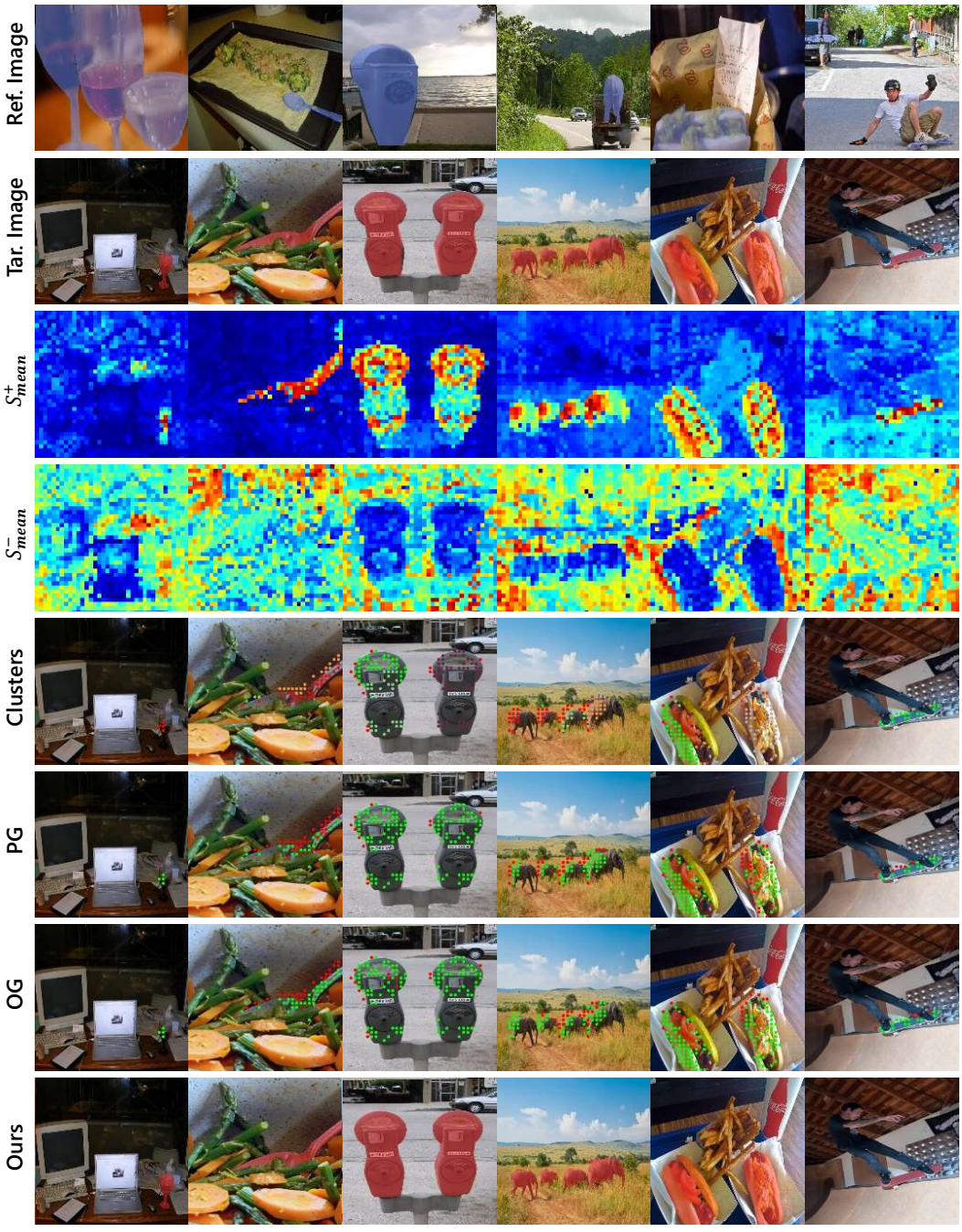


Figure 12: Qualitative analysis of the contents in gating. Different colors of points in the images in column “Clusters" represent different clusters. The green points in images in columns “PG" and “OG" denote the points satisfying the gating criteria, while the red points denote those not satisfying.

https://doi.org/10.52202/079017-1047

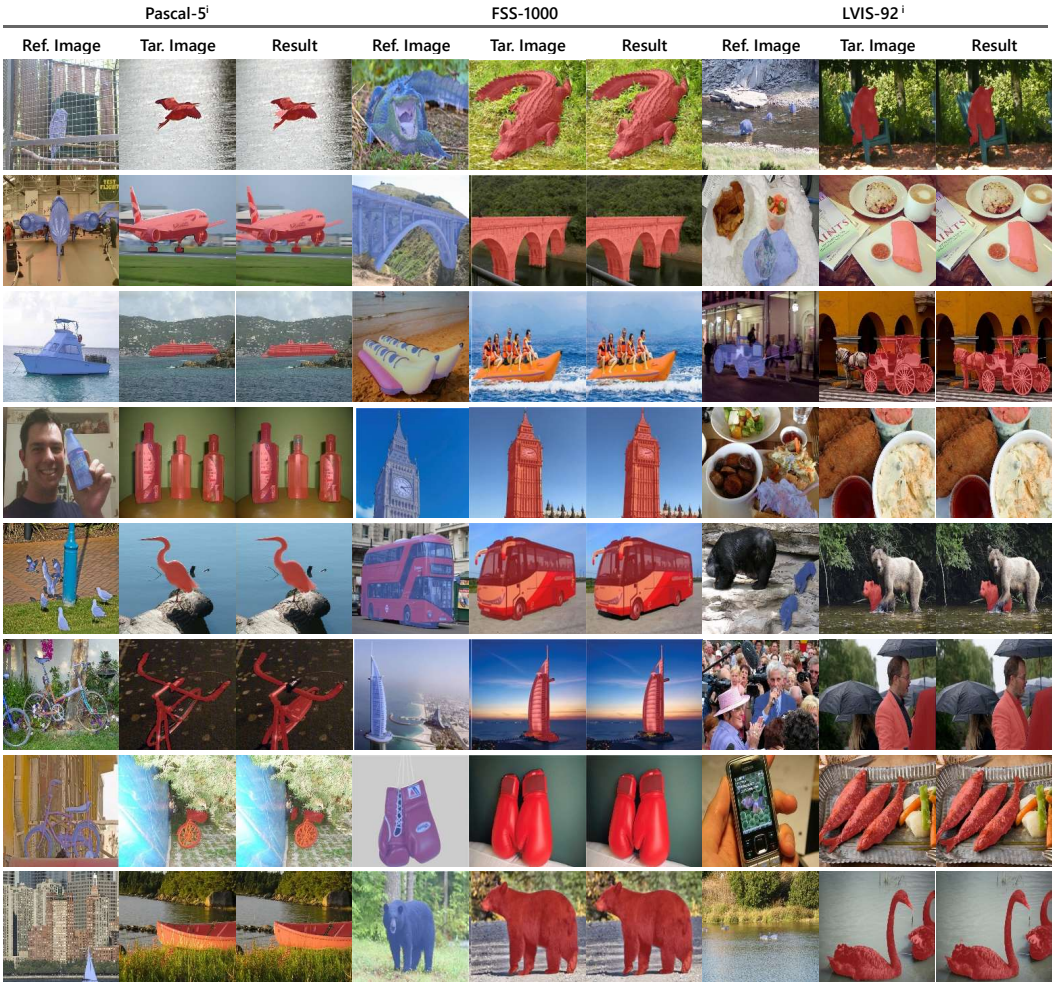


Figure 13: Qualitative analysis of the results on Pascal-5^i, FSS-1000, and LVIS-92^i.

https://doi.org/10.52202/079017-1047

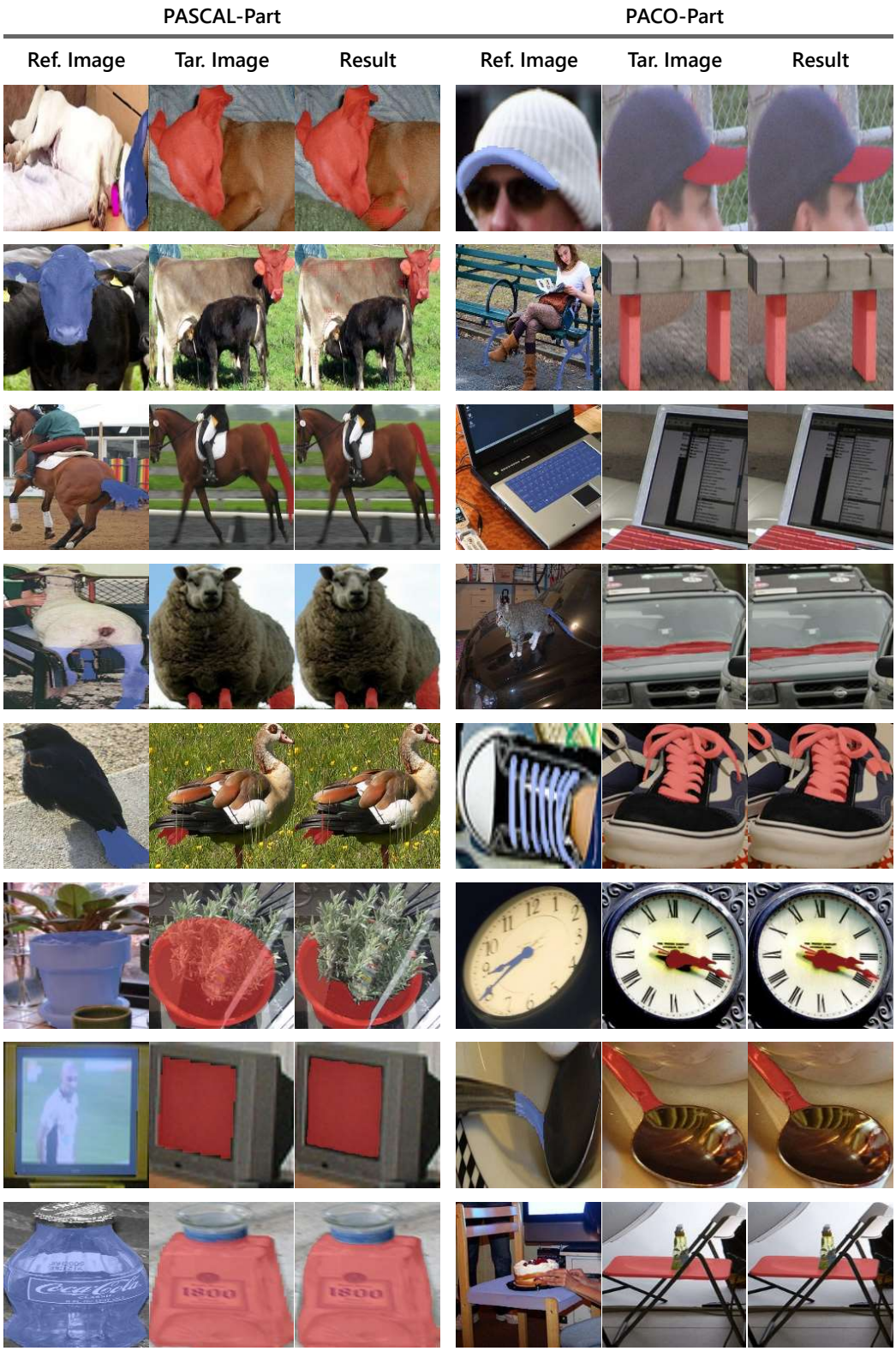


Figure 14: Qualitative analysis of the results on PASCAL-Part and PACO-Part.

https://doi.org/10.52202/079017-1047

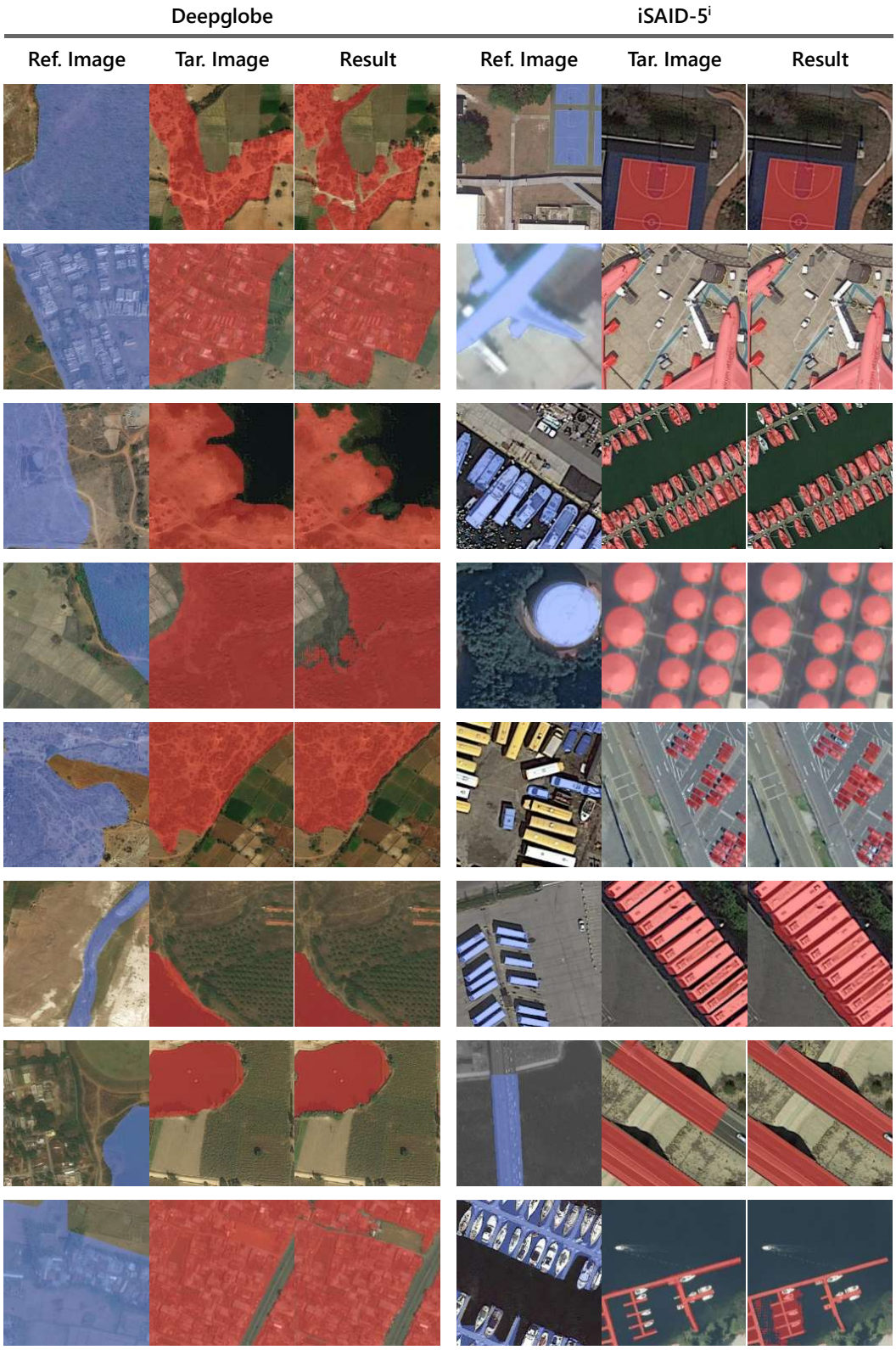


Figure 15: Qualitative analysis of the results on Deepglobe and iSAID-5^i.

https://doi.org/10.52202/079017-1047

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: Our main claims made in the abstract and the last two paragraphs of the introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA 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 NA 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: We discuss the limitations in the appendix.

   Guidelines:

   - The answer NA 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: [NA]

https://doi.org/10.52202/079017-1047

Justification: We do not introduce theory assumptions and proofs.

Guidelines:

- The answer NA 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: Our paper fully discloses all the information needed to reproduce the main experimental results of the paper.

   Guidelines:

   - The answer NA 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: [Yes]

Justification: Our code and instructions are included in the supplementary material. The data we use for the experiments are all from open-access datasets.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Our paper specifies all the test details for our training-free approach in the section of experiments.

   Guidelines:

   - The answer NA 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: We set 5 random seeds to evaluate the large datasets and evaluate all samples of the small datasets in the appendix.

   Guidelines:

   - The answer NA 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.)

https://doi.org/10.52202/079017-1047

- 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: Computer resources are described in Implementation Details of the experiment section.

   Guidelines:

   - The answer NA 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: Our research conforms with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA 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: We discuss both potential positive societal impacts and negative societal impacts of our work in the appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper has no such risks.

    Guidelines:

    - The answer NA 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: The existing assets used in our paper, i.e., SAM and DINOv2, are released on GitHub under Apache License 2.0.

    Guidelines:

    - The answer NA 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.

https://doi.org/10.52202/079017-1047

- 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: [Yes]

    Justification: Our new assets introduced in the paper are well documented in the supplementary material.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: There is no crowdsourcing and research with human subjects in our paper.

    Guidelines:

    - The answer NA 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: [NA]

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

    Guidelines:

    - The answer NA 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.