Training-free Boosting for Few-shot Segmentation via Generalizing Semantic Mining

Kangyu Xiao[1], Zilei Wang [1 *], Yixin Zhang[2], Junjie Li[3]
[1]University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[3]Beijing University of Posts and Telecommunications
xiaoky@mail.ustc.edu.cn, {zlwang, zhyx12}@ustc.edu.cn, hnljj@bupt.edu.cn

Abstract

Few-shot semantic segmentation (FSS) aims to segment the novel target objects with the guidance of minimal annotated reference examples. The affinity-based method has great advantages in the FSS inference stage for both specialist model and foundation model. However, current affinity calculation merely relies on only support-query matching, without considering the query-specific semantic or the semantic correlation among inter-support samples, which limits the representation ability of affinity map. In this paper, we propose the Generalizing Semantic Mining (GSM) that focuses on exploiting generalizing semantic to improve the affinity calculation. Concretely, we first organize the affinity-based inference into three main steps to reveal the crucial role of affinity map. To address the low-data problem, Target Semantic Reusing module considers the query sample as a proxy reference and assigns it with proxy mask identifying its most generalizing semantic regions. Then, to generate the high-fidelity proxy mask, Query-specific Semantic Modeling module pinpoints the most generalizing regions through prior semantic analysis. Finally, Representative Re-weighting module explicitly modulates affinity calculation via generalization-aware weighting. Experiments on FSS benchmarks demonstrate that our GSM can serve as a plug-and-play free lunch for both specialist models and foundation models.

Introduction

Few-shot semantic segmentation (FSS) aims to semantically segment the novel target objects with the guidance of minimal annotated reference examples. To achieve this, current approaches design a support-query knowledge propagation mechanism, and they can be roughly divided into two categories, namely prototype-based method and affinity-based method. Prototype-based methods (Wang 2024a; Park et al. 2025; Li et al. 2021) summarize the support objects into global prototypes for query features similarity comparison. Affinity-based methods (Park et al. 2024; Shi et al. 2022a; Wang, Sun, and Zhang 2023) leverage dense feature for pixel-level correlation (affinity map) between support and query. Comparing prototype-based method with affinity-based method, the former risks spatial information loss during prototype summarization. Besides, current visual foundation model (*e.g.,* SAM (Kirillov et al. 2023)) can provide impressive dense prediction based on the given visual prompts, which is highly flexible. Since affinity-based method emphasizes pixel-wise matching, it can naturally select representative point prompts and perfectly apply to the visual reference prediction of foundation models. Following the above advantages, our method is also affinity-based and focuses on the FSS inference stage of the specialist model and foundation segmentation model (SAM). We organize the affinity-based inference stage into three main steps, as shown in Figure 1. Clearly, whether for specialist model or foundation model, affinity map is always crucial.

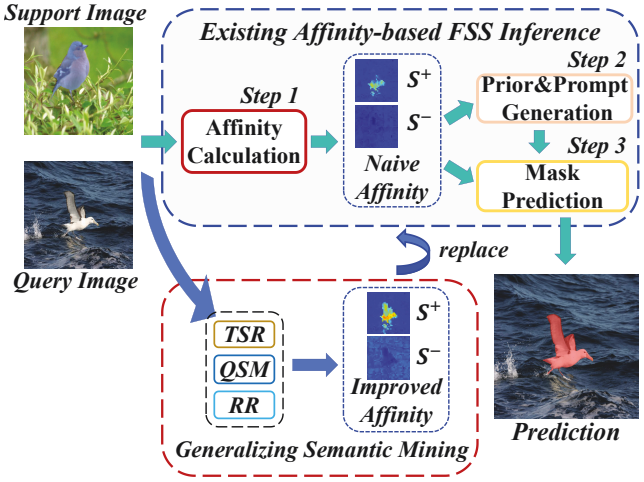

Figure 1: An illustration of affinity-based FSS inference.We can see that the affinity map closely guides the step 2&3, and our GSM is a plug-and-play boosting for affinity calculation.

However, current affinity calculation merely relies on only support-query matching, without considering the query-specific prior semantic or the semantic correlation among inter-support samples, limiting the representation ability of affinity map. Particularly, **for query image**, the object regions possess spatial continuity, *e.g.,* a foreground pixel implies high likelihood of foreground membership in its neighborhood. While current methods fail to explicitly model such adjacency relationship, which fragment the se-

*the corresponding author.

(a) (b) (c)

Figure 2: A comparison of samples' generalizing capacity. The semantic category is **person**. (a) and (c) generate structurally coherent semantic representation, while (b) produces overly coarse semantic.

mantic continuity in affinity map and result in spatially incoherent pixels. Such semantically deficient affinity map will directly affect the prior or prompt guided mask prediction. **For reference set**, as shown in Figure 2, when there are multiple support samples (*e.g.,* 3-shot FSS), it always exists a sample with stronger generalizing capacity that can serves as a more generalizing reference. Conversely, there may also exists a sample containing ambiguous information, thereby harming the affinity calculation. While most methods ignore such semantic differences and just directly take the sample-wise mean of affinity maps. Besides, the **low-data problem** is also challenging for affinity calculation.

In this paper, we aim to improve the affinity calculation in a training-free manner to provide a free lunch for affinity-based methods. The core insights include: 1) A training-free design can eliminate the computational burden in real-world application, which especially preserves the out-of-box usability and rich prior knowledge of prompt-able foundational models; 2) For reference set, the most generalizing components should be highlighted, while the low quality components should be weakened; 3) The query sample can enrich the reference knowledge, and we can find the high-fidelity proxy mask by exploring query-specific semantic.

To this end, we propose Generalizing Semantic Mining (GSM), a novel training-free boosting for affinity-based FSS inference, including Target Semantic Reusing (TSR), Query-specific Semantic Modeling (QSM) and Representative Re-weighting (RR). Particularly, we first organize the affinity-based inference into three main steps to reveal the importance of improving the affinity map. Then, to address the low-data problem, we develop TSR module to enrich the reference knowledge with query sample and assign it with proxy mask identifying query's most generalizing semantic regions. To generate the high-fidelity proxy mask, our QSM module pinpoints the most generalizing regions through prior semantic analysis. Finally, our RR module explicitly modulates affinity calculation via generalization-aware weighting. We conduct experiments on various datasets, including in domain and out of distribution scenarios, which reveals that our proposed GSM can serve as a plug-and-play free lunch for both specialist models and foundation models.

The contributions are summarized as follows.

- We summarize the affinity-based FSS inference into three steps, and discuss the critical role of affinity map.
- We propose GSM, a training-free boosting for affinity-based FSS inference, in which TSR leverages query sample to enrich the reference knowledge, QSM is to generate the high-fidelity proxy mask, and RR is to highlight the generalizing support component.
- We conduct extensive experiments on FSS benchmarks to reveal that GSM can serve as a plug-and-play free lunch for both specialist models and foundation models.

Related Work

Few-shot Semantic Segmentation

Few-shot semantic segmentation (FSS) aims to segment target objects with the guidance of minimal annotated reference examples. Predominant approaches fall into two categories, namely prototype-based methods and affinity-based methods. Prototype-based methods (Wang 2024a; Park et al. 2025; Li et al. 2021) summarize the support objects into global prototypes for query features similarity comparison. Affinity-based methods (Park et al. 2024; Shi et al. 2022a; Wang, Sun, and Zhang 2023) leverage dense feature for pixel-level correlations between all support and query. Comparing prototype-based method with affinity-based method, the former risks spatial information loss during prototype summarization. Moreover, current visual foundation models support multiple visual prompts. For example, SAM (Kirillov et al. 2023) and DINOv (Li et al. 2024) can provide dense prediction based on given prompts (Li et al. 2025), which is highly flexible. Since affinity-based method emphasizes pixel-wise matching, it can naturally select representative points and perfectly apply to the visual reference prediction of foundation models. Following such advantages, our method is also affinity-based and focuses on the visual reference task (*i.e.,* FSS) based on specialist models and foundation segmentation model SAM.

SAM-based Visual Reference Segmentation

The recent Segment Anything Model (SAM) (Kirillov et al. 2023) can exhibit exceptional prompt-able zero-shot segmentation ability based on given visual prompts like point, box and coarse mask. SAM has natural advantage for visual reference tasks like FSS, since the support-query knowledge transduction can easily extract visual prompts. Existing SAM-based FSS methods commonly leverage pixel-wise matching to obtain SAM's visual prompt, call SAM to generate coarse-grained segmentation, and then post-process it to obtain the final prediction mask. PerSAM (Zhang 2024b) first employs SAM for prompt-able segmentation with one-shot guidance. Matcher (Liu 2024) and GF-SAM (Zhang 2024a) propose simple and highly efficient training-free inference baselines through prompt refinement and graph analysis. The inference stage can be organized into three main steps: affinity calculation, prior & prompt generation, mask prediction. Importantly, the core steps of visual reference (*i.e.,* prior & prompt generation, mask prediction) are closely guided by the affinity map.

Method

Problem Formulation

Few-shot semantic segmentation (FSS) aims to segment target objects with the guidance of minimal annotated refer-

ence examples. We focus on boosting the inference stage. During inference, it is given two groups of samples. The support reference set $S = \{x_n^s, m_n^s\}_{n=1}^k$ contains image-mask pairs, where x_n^s is the n-th support image sample, m_n^s is the corresponding binary foreground (FG) mask, $\neg m_n^s$ reveals the background (BG) mask, and k is the number of annotated pairs, which is defined as k-shot segmentation problem. The query inference image x^q is needed to be semantically segmented according to S. Here $x^s \in \mathbb{R}^{H\times W\times C}$, $x^q \in \mathbb{R}^{H\times W\times C}$, and $m^s \in \mathbb{R}^{H\times W}$.

Method Overview

In this work, we first review the affinity-based FSS inference and summarize it into three steps to discuss the critical role of affinity map. Then we propose Generalizing Semantic Mining (GSM), a training-free boosting for affinity-based inference, including Target Semantic Reusing (TSR), Query-specific Semantic Modeling (QSM) and Representative Re-weighting (RR). Figure 1 shows our plug-and-play pipeline, and Figure 3, 4 illustrate our proposed modules.

A Review of Affinity-based Method

Affinity-based methods leverage pixel-level correlations between x^s and x^q. Here we review with the GF-SAM, since it constructs a universal baseline for affinity-based and SAM-based FSS. We can summarize the calculation and application of affinity into 3 steps. We can see that, **the core steps of visual reference (*i.e.,* prior & prompt generation, mask prediction) are closely guided by the affinity map.**

Step 1: Affinity calculation. A backbone f extracts dense features as F^s and $F^q \in \mathbb{R}^{h\times w\times c}$. The pixel-wise similarity between F^s and F^q is calculated as $C(i,j) = cos\_sim(F^q(i), F^s(j))$, $C \in \mathbb{R}^{hw\times hw}$, where $C(i,j)$ reveals the similarity between the i-th pixel of query $F^q(i)$ and the j-th pixel of support $F^s(j)$. With the m^s, C can be split to C^+ and $C^- \in \mathbb{R}^{h\times w}$. C^+ is the similarity between F^s FG region and each F^q pixel. While C^- is the similarity between F^s BG region and each F^q pixel. For k-shot support examples, it gets the positive and negative affinity sets as $A^+ = \{C_n^+\}_{n=1}^k$ and $A^- = \{C_n^-\}_{n=1}^k$. For existing affinity-based methods like PerSAM and GF-SAM, when $k \geq 2$, they **calculate the mean of affinity** like $S^+ = \frac{1}{k}\sum A^+$ and $S^- = \frac{1}{k}\sum A^-$. S^+ and $S^- \in \mathbb{R}^{h\times w}$.

Step 2: Prior & prompt generation. The prior mask is considered as $m_{prior} = S^+ > S^-$, and the point, mask or box prompts of SAM will be generated according to it.

Step 3: Mask prediction. SAM generates the coarse-grained masks through prompts. The post-processing will filter and aggregate the candidate masks by C, S^+ and S^-, then give the final predicted mask m_{pred}. For more detailed process, please refer to the specific methods, like GF-SAM.

Target Semantic Reusing

For few-shot learning, the low-data problem is always fatal (Xiao et al. 2024). A straightforward approach involves enriching reference knowledge using query data. While the core challenge lies in designing effective proxy labels. We first define a new reference set as,

$$S^* = \{(x_1^s, m_1^s), ..., (x_k^s, m_k^s), (x^q, m_{proxy})\}, \quad (1)$$

here $m_{proxy} = (m_{pos}, m_{neg})$, representing the most generalizing FG and BG semantic regions of x^q, and they are not required to be strictly complementary.

Obviously, the best candidate for m_{proxy} must be the ground-truth mask m_{gt}^q of x^q. This implies setting $m_{pos} = m_{gt}^q$ and $m_{neg} = \neg m_{gt}^q$. In practice, however, we are unable to access m_{gt}^q. The second candidate is m_{pred}, but it clearly has many ambiguous semantics due to the constraints of baseline methods and data distribution shifts.

Specifically, for m_{pos}, false positive semantics are particularly detrimental: they misguide the model into recognizing BG regions as highly generalizing FG regions. In contrast, false negative semantics pose a secondary concern, as the omitted FG regions typically lack generalizing power and contribute less to affinity calculation. In summary, a high true positive rate in m_{pos} is critical for confidently highlighting generalizing FG regions. The same applies to m_{neg}.

Query-specific Semantic Modeling

QSM derives high-quality proxy mask by leveraging query-specific semantics. Particularly, there are some image-own natural priors that can enhance the reliability of the proxy mask. We summarize these priors as 1) feature similarity, 2) pixel-level spatial similarity, 3) region-level spatial similarity. Current methods fail to explicitly exploit such priors, leading to sub-optimal affinity and mask representations. While we explicitly model them to obtain pixel relationship descriptors, and employ clustering algorithms to get high generalization regions. The pipeline is shown in Figure 3.

Prior 1: feature similarity modeling Obviously, the higher the feature similarity between pixels, the more likely they are to belong to the same object. To quantify this relationship, we calculate the pixel-wise correlation,

$$C_{self} = \frac{F^q \cdot (F^q)^\top}{\|F^q\|^2}, \quad (2)$$

here $C_{self}(i,j)$ indicates the feature similarity between the i-th and j-th pixel, $C_{self} \in \mathbb{R}^{hw\times hw}$.

Prior 2: pixel-wise spatial similarity modeling A FG pixel implies high likelihood of FG membership in its neighborhood. We use a Gaussian kernel to model the continuity relationship between neighboring pixels,

$$K(i,j) = \exp\left(-\frac{\|i-j\|^2}{2\sigma^2}\right) \quad (3)$$

here $K(i,j)$ indicates the pixel-wise spatial similarity between the i-th and j-th pixel, the proximity closer i and j, the higher the $K(i,j)$, and $K \in \mathbb{R}^{hw\times hw}$.

Prior 3: region-level spatial similarity modeling Object itself has local region similarity. For two regions, a high region similarity means they are likely to represent the same type of object. For example, a cat's two ears have similar

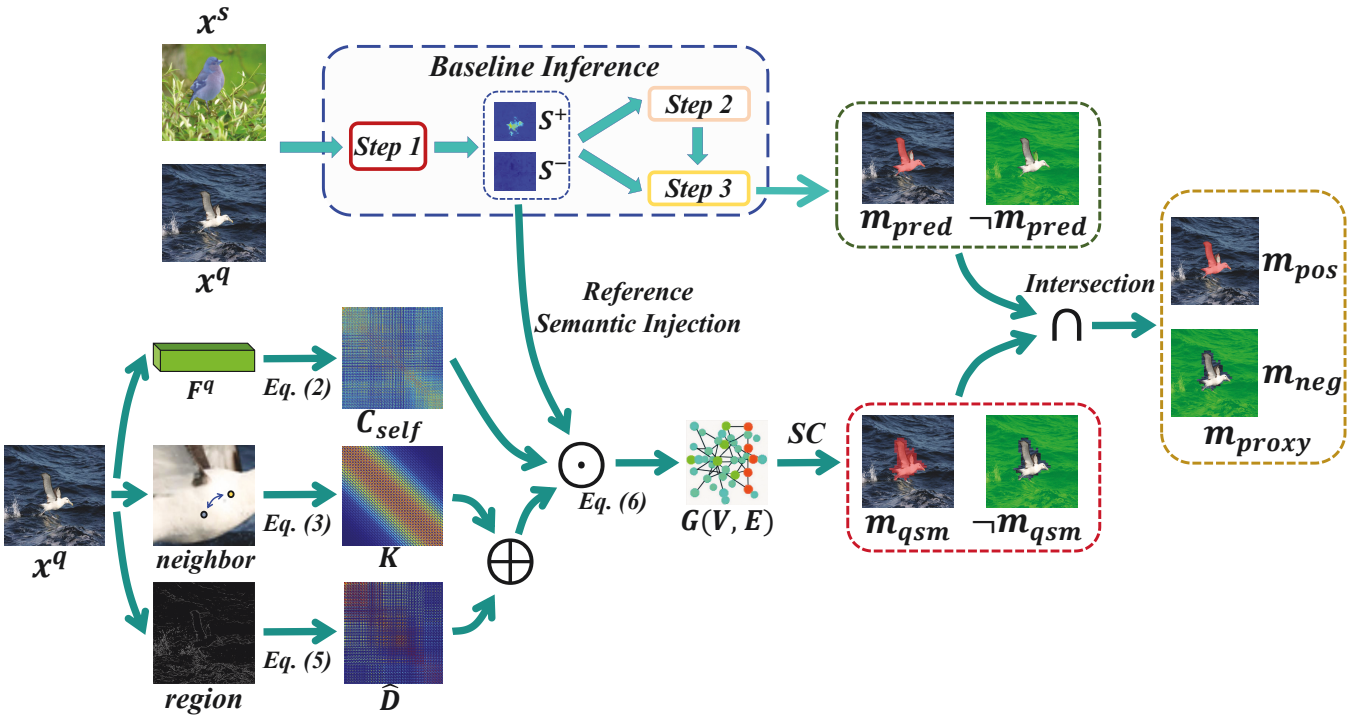

Figure 3: Illustration of Query-specific Semantic Modeling. We model pixel relationships by query-own natural priors to identify the most generalizing semantic regions. Here SC represents the spectral clustering.

shapes. So we can use region-level semantics to describe the query-own spatial similarity. We first use an edge operator to perform coarse-grained region splitting for x^q,

$$R = \{R_1, R_2, ..., R_l\} = \mathcal{F}(\mathcal{I}(edge(x^q))). \quad (4)$$

Here, $edge$ is an edge operator, $\mathcal{I}$ is an interpolation to $h \times w$, and $\mathcal{F}$ is the flattening operation. We thus split the $h \times w$ pixels of F^q into l spatially contiguous regions.

We then analyze the inter-region correlations to adaptively refine them via splitting or merging. Particularly, the region similarity is computed as the mean correlation across all inter-region pixel pairs,

$$D_{p,q} = \frac{1}{|R_p| \cdot |R_q|} \sum_{i \in R_p} \sum_{j \in R_q} C_{self}(i, j), \quad (5)$$

here $D \in \mathbb{R}^{l \times l}$. We can assign the region's descriptor D to all constituent pixels, generating a region affinity map $\hat{D} \in \mathbb{R}^{hw \times hw}$ that encodes pixel-wise spatial similarity at the region level (*i.e.,* region-level spatial similarity).

Graph analysis Through the modeling, we get three types of relationships between query pixels, *i.e.,* feature level, neighbor level and region level. Our goal is to use these descriptors to split out the most generalizing FG and BG regions. Given the known relationship, the most effective way for dividing pixels is spectral clustering (Ng et al. 2001).

Spectral clustering is based on graph splitting, so we need to model the pixel relationships as a directed graph $G(V, E)$, where V denotes the $h \times w$ pixels and E represents directed edges encoding pairwise pixel semantic affinities. We want to obtain an edge that pixels with higher feature similarity, closer distance, and sharing higher region similarity will be confidently described into a new semantically dense region. Therefore, we define the edge relationship as a joint representation coupling feature similarity and spatial similarity,

$$E = C_{self} \odot (K + \hat{D}) \odot \mathcal{B}(\mathcal{F}(\exp\left(S^+ - S^-\right))). \quad (6)$$

Here, C_{self} is the feature similarity, $(K + \hat{D})$ indicates the spatial similarity, $\odot$ is the Hadamard product, $\mathcal{B}$ is the linewise broadcast for dimension matching, $(S^+ - S^-)$ introduces the semantic affinity between query pixels and reference information, and exp is to amplify the reference semantics. It should be noted that, K is a short-distance spatial prior, while $\hat{D}$ is a long-distance spatial prior, they complement each other in terms of scale. If either of them has large value, it indicates a high spatial similarity between pixels. So their coupling adopts union, *i.e.,* adding. For feature similarity, spatial relationship, reference information, they have synergistic effect. Only when all of them are large can a pixel belong to target object. So we describe them as intersection, *i.e.,* product. E is a comprehensive consideration that combines query-own and target-guided semantic. A high-weight edge $E(i, j)$ indicates a correlation congruent node pair $i-j$ that conforms to support-induced target semantic.

By splitting the graph, we can aggregate spatially contiguous pixels exhibiting semantic consistency in both local affinity and reference alignment. We then employ spectral clustering to perform fine-grained splitting,

$$R^* = \{R_1^*, R_2^*, ..., R_m^*\} = \mathcal{SC}(G(V, E)), \quad (7)$$

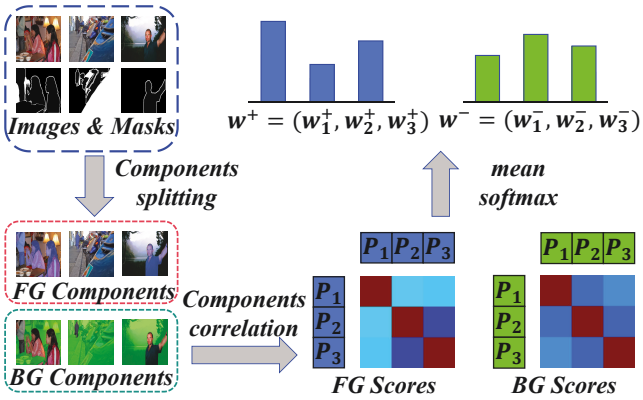

Figure 4: Illustration of Representative Re-weighting. We split images into FG&BG components, calculate intra component correlation, get the generalization score as weight.

here $\mathcal{SC}$ represents spectral clustering.

Semantic aggregation Through graph analysis, we assign fine-grained regional labels to each pixel and consolidate representative FG pixels into coherent semantic groupings. Based on the more detailed region information, we can reuse Eq. 5 & 6 to update the region similarity and get new edge weights $\overline{E}$, which is then subjected to a binary clustering to obtain the binary mask m_{qsm}.

$$\{m_{qsm}, \neg m_{qsm}\} = \mathcal{SC}(G(V, \overline{E})) \tag{8}$$

While m_{qsm} captures the region with generalizing semantics, it may exhibits low confidence areas, especially at the boundaries of target objects. This occurs because, in boundary areas, low feature similarity conflicts with high Gaussian weighting, leading to sub-optimal clustering and segmented boundaries. To filter the low confidence areas, we integrate the predictions from both m_{qsm} and m_{pred},

$$\begin{cases} m_{pos} = m_{qsm} \cap m_{pred} \\ m_{neg} = \neg m_{qsm} \cap \neg m_{pred}. \end{cases} \tag{9}$$

Specifically, when their predictions are consistent, we can infer with high confidence that a region exhibits the most generalizing semantics.

Representative Re-weighting

For reference set, the most generalizing components should be highlighted, while the low quality components should be weakened. From this perspective, RR quantifies the generalization capability of FG and BG components among support examples, assigning them adaptive weights for affinity calculation. The pipeline is shown in Figure 4.

Intra-component correlation. A sample with strong generalizing capacity will exhibit generally high similarity with other samples. So we measure intra-component similarity (*e.g.,* FG-to-FG & BG-to-BG) across samples to quantify it.

Particularly, we first calculate the FG and BG prototypes of each sample $x \to (P_{fg}, P_{bg})$ as,

$$\begin{cases} P_{fg} = \frac{\sum_1^{h\times w} F^s \odot \mathcal{I}(m^s)}{\sum_1^{h\times w} \mathcal{I}(m^s)} \\ P_{bg} = \frac{\sum_1^{h\times w} F^s \odot \mathcal{I}(\neg m^s)}{\sum_1^{h\times w} \mathcal{I}(\neg m^s)} \end{cases} \tag{10}$$

For the proxy sample (x^q, m_{proxy}) from TSR, we use m_{pos} instead of m^s and m_{neg} instead of $\neg m^s$. We thus obtain the support prototypes set $P = \{(P_{n,fg}, P_{n,bg})\}_{n=1}^{k+1}$.

Then, we calculate the intra-component similarity as

$$C_{fg}(i,j) = cos\_sim(P_{i,fg}, P_{j,fg}), \tag{11}$$

indicating the similarity between the FG components of the i-th and j-th sample. The calculation for BG is in the same way, and $C_{fg}, C_{bg} \in \mathbb{R}^{(k+1)\times(k+1)}$.

Generalization quantification. To quantify the overall similarity of a sample to the other samples, we calculate the mean of C_{fg} by column as,

$$w_{i,fg} = \frac{1}{k} \sum\nolimits_{j=1, j\neq i}^{k+1} C_{fg}(i,j). \tag{12}$$

Evidently, a larger w_i signifies that the i-th sample exhibits stronger generalizing representation of the target object. By contrary, a small w_i suggests the presence of ambiguous semantics and is not competent as a general representation.

Once quantifying the generalizing capacity of components, we perform $softmax$ to obtain their corresponding weights,

$$w^+ = softmax(w_{1,fg}, \ldots, w_{k+1,fg}). \tag{13}$$

Then we can apply this weight in affinity calculation,

$$\overline{S}^+ = \frac{1}{k+1} \sum\nolimits_{j=1}^{k+1} w_i^+ C_i^+. \tag{14}$$

The calculation for BG components is in the same way.

We finally get the re-weighted affinity map $\overline{S}^+$ and $\overline{S}^-$, which can be used for refining the prior & prompt generation and mask prediction step of the affinity-based inference.

Experiments

Experimental Settings

Datasets We conduct experiments on in-domain FSS (ID-FSS) and cross-domain FSS (CD-FSS) datasets. The **ID-FSS** datasets including Pascal-5^i, COCO-20^i and LVIS-92^i. The **CD-FSS** datasets including Deepglobe, ISIC2018, iSAID-5^i and FSS-1000. Details please refer to appendix.

Implementation details To validate that GSM is a free lunch for affinity-based methods, we equip GSM on existing specialist (TBS) and foundation (Matcher & GF-SAM) baselines on ID-FSS and CD-FSS settings. For fair comparison, the inference stage is strictly followed the baseline methods. Particularly, when equipping GSM on TBS (Park et al. 2024), we use Swin-Transformer as feature extractor f, and DCAMA (Shi et al. 2022b) for mask prediction. When equipping GSM on Matcher (Liu 2024) and GF-SAM (Zhang 2024a), we use DINOv2 (Oquab et al. 2023) with ViT-L/14 as as feature extractor f, and SAM with ViT-H as mask generator. We report the mean intersection over union (mIoU) for performance evaluation. All experiments are conducted on a single NVIDIA RTX 3090.

In-domain FSS	Pascal-5^i		COCO-20^i		LVIS-92^i	
Methods	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
specialist models						
HSNet(Min 2021)	66.2	70.4	41.2	49.5	17.4	22.9
VAT(Hong et al. 2022)	67.9	72.0	41.3	47.9	18.5	22.7
HDMNet(Peng 2023)	69.4	71.8	50.0	56.0	-	-
AMFormer(Wang 2023)	70.7	73.6	51.0	57.3	-	-
ABCB(Zhu et al. 2024)	72.0	74.9	51.5	58.8	-	-
AENet(Xu et al. 2024)	70.3	74.2	51.3	57.1	-	-
Ada-FSS(Wang 2024a)	72.3	79.1	48.3	60.0	-	-
TBS(Park et al. 2024)	71.2	75.2	52.3	58.5	-	-
TBS+GSM (ours)	74.4	77.6	54.2	61.9	-	-
foundation models						
PerSAM(Zhang 2024b)	43.1	-	23.0	-	11.5	-
PI-CLIP(Wang 2024b)	**76.8**	77.2	56.8	59.1	-	-
FCP(Park et al. 2025)	73.2	74.0	52.5	58.0	-	-
Matcher(Liu 2024)	68.1	74.0	52.7	60.7	33.0	40.0
Matcher+GSM (ours)	71.3	75.7	55.9	62.7	34.3	41.9
GFSAM(Zhang 2024a)	72.1	82.6	58.7	66.8	35.2	44.2
GFSAM+GSM (ours)	75.7	**83.3**	**61.2**	**68.1**	**38.9**	**47.0**

Table 1: Comparison (mIoU) on ID-FSS datasets. We both consider methods for specialist and foundation models.

Main Results

Comparison on In-domain FSS We compare GSM with SOTA specialist and foundation methods on ID-FSS datasets in Table 1. Notably, both the specialist model TBS and the foundation models Matcher & GF-SAM achieve significant gains when boosted by GSM. Particularly, GF-SAM + GSM establishes new SOTA results on most datasets. This demonstrates the strong plug-and-play capability and generality of GSM as a performance booster for FSS.

From the results, we also have the following observations. 1) Reference Number. As expected, performance generally improves with more shots, confirming that additional reference information enhance affinity calculation. That's why we recommend enriching the reference knowledge through query sample. 2) Foundation vs. Specialist. Recent methods leveraging foundation models (*e.g.,* GF-SAM, PI-CLIP) consistently outperform specialist models designed solely for FSS (*e.g.,* TBS, ABCB), underscoring the advantage of transferring rich, general-purpose visual prior knowledge. Though current specialist models show incremental progress, their performance remains below that of leading foundation model based methods. 3) Dataset Difficulty. Performance varies significantly across datasets, highlighting their relative difficulty. For example, LVIS-92^i proves to be a challenging benchmark, as evidenced by considerably lower absolute mIoU across all methods. This is because it suffers from severe fine-grained and long-tail challenges.

Comparison on Cross-domain FSS We show comparison on CD-FSS datasets in Table 2. Results are similar to ID-FSS, models achieve significant gain when boosted by GSM, and also establish new SOTA results on most datasets.

Ablation Study

We analyze the module-level effectiveness and computational cost to show GSM is a plug-and-play free lunch for FSS inference. For more ablation, please refer to appendix.

Component Ablation GSM has three modules, *i.e.,* TSR, QSM, RR. To validate their contribution, we conduct components ablation. We report the average mIoU on the ID-FSS (3 datasets) and CD-FSS (4 datasets) settings, in Table 3. It should be noted that QSM can only be used together with TSR; and when TSR is used alone, we use m_{pred} as m_{proxy}.

It can be observed that, 1) Using TSR without QSM yields marginal gains (A vs. B, D vs. E). Though additional reference data holds potential, the ambiguous semantics in m_{pred} disturb the reference information. 2) Combining TSR and QSM can achieve substantial gains (A vs. C, D vs. F), confirming that mining generalizing query semantic is helpful. Gains are particularly pronounced in 1-shot settings, since the additional reference information effectively mitigates the low-data problem. 3) RR is invalid for 1-shot, because when there is only one or two samples, it cannot to obtain a more generalizing sample through comparison. But when there are more samples (*e.g.,* 5-shot), it can always bring gains by generalization-aware re-weighting. In summary, the combination of the above three modules, *i.e.,* GSM, consistently boosts FSS performance across diverse few-shot conditions.

Computational Cost To demonstrate that GSM is a plug-and-play free lunch for FSS inference, we compare the current SAM-based methods on inference speed, GPU memory requirement and FSS performance, in Table 4. Though PerSAM offers the fastest inference (1.43s) and lowest memory (5.8 GB), its accuracy is so low. Remarkably, our substantial gains are achieved with minimal computational cost. Integrating GSM increases inference time by only 0.25s per

Cross-domain FSS	Deepglobe		ISIC		iSAID-5^i		FSS-1000	
Methods	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
specialist models								
HSNet(Min 2021)	29.7	35.1	31.2	35.1	34.1	40.4	86.5	88.5
ABCD(Herzog 2024)	42.6	49.0	45.7	53.3	-	-	74.6	76.2
DRAdapter(Su 2024)	41.3	50.1	40.8	48.9	-	-	79.1	80.4
foundation models								
PerSAM	31.4	-	23.9	-	19.2	-	71.2	-
APSeg(He 2024)	35.9	40.0	45.4	54.0	-	-	79.7	81.9
Matcher(Liu 2024)	48.1	50.9	38.6	35.0	33.3	34.3	87.0	**89.6**
Matcher+GSM	50.6	54.2	42.7	44.5	36.8	39.1	87.3	89.6
GFSAM	49.5	57.7	48.7	55.2	47.1	52.4	88.0	88.9
GFSAM+GSM	**51.4**	**58.9**	**52.2**	**58.6**	**49.8**	**53.9**	**88.6**	89.3

Table 2: Comparison (mIoU) on CD-FSS datasets.

	Components			In-domain FSS		Cross-domain FSS	
Index	TSR	QSM	RR	1-shot	5-shot	1-shot	5-shot
A(GFSAM)				55.3	64.5	58.3	63.5
B	√			56.5	64.2	58.5	63.9
C	√	√		58.6	65.7	60.5	64.8
D			√	55.3	65.3	58.3	64.1
E	√		√	56.5	65.1	58.5	64.4
F	√	√	√	58.6	66.1	60.5	65.2

Table 3: Component Ablation. Gain of employing GSM.

Methods	PerSAM	PerSAM-F	Matcher	GFSAM	GSM
Training-free	√		√	√	√
Speed (s/img)	1.43	16.5	12.7	1.88	2.13
GPU memory	5.8 GB	5.8 GB	8 GB	8.9 GB	9.3 GB
ID-FSS mIOU	25.9/-	28.1/-	51.3/58.2	55.3/64.5	58.6/66.1
CD-FSS mIOU	36.4/-	38.6/-	51.8/52.5	58.3/63.5	60.5/65.2

Table 4: Comparison of Computational Cost.

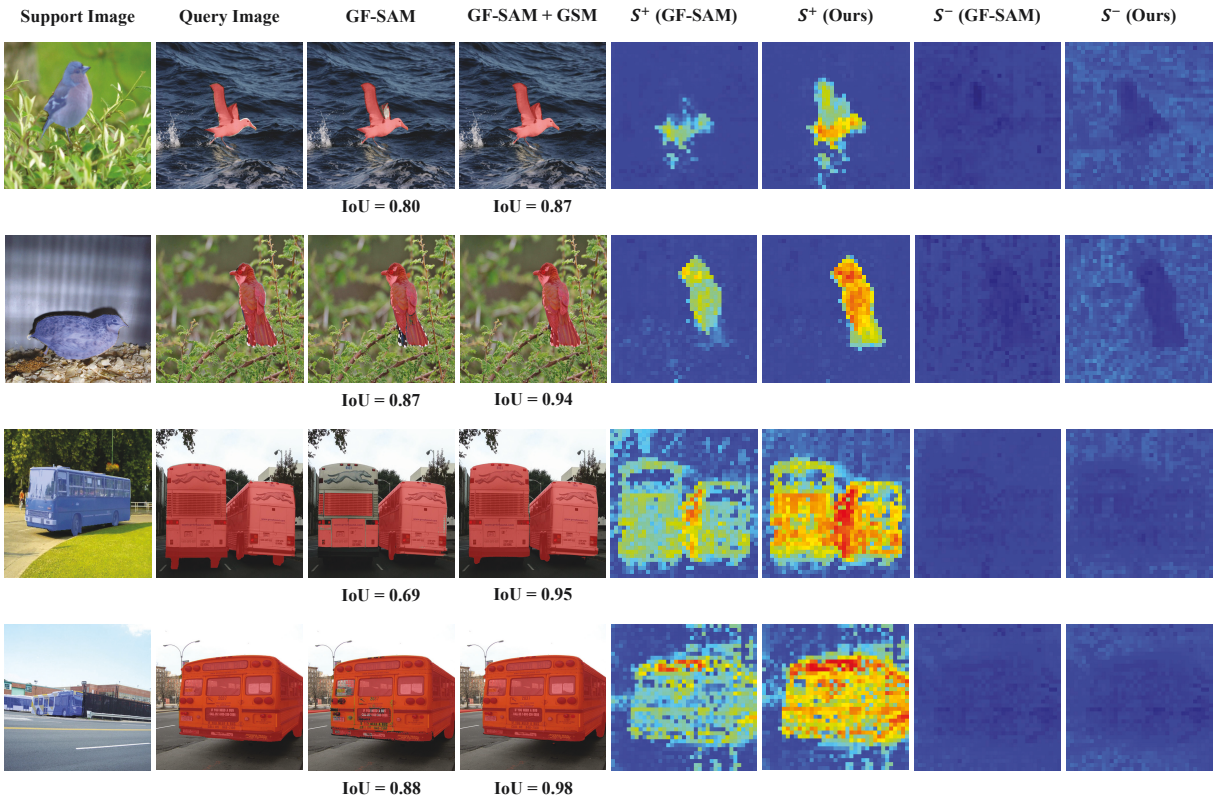

Figure 5: Visualization analysis of GF-SAM and GF-SAM + GSM (Ours). We compare the predictions and affinity maps of GF-SAM and Ours. Here, warmer colors reveal higher values. Masks of support and query are shown in blue and red, respectively.

image (1.88s $\rightarrow$ 2.13s) and GPU memory requirement by 0.4 GB (8.9 GB $\rightarrow$ 9.3 GB), while maintaining the essential training-free property. Critically, the training-free design preserves foundation model's flexibility and rich prior knowledge, mitigating overfitting risk in low-data situations.

It should be noted that, though GSM executes the baseline inference twice, the runtime does not be doubled. This is because SAM-based methods incur most cost from large-model calling (*e.g.,* DINOv2 and SAM). And our boosted affinity map only supplements sparse high-confidence point prompts and then refines the post-processing stage, thereby adding minimal SAM calling and computation cost.

Visualization analysis We present the visualization results in Figure 5. Besides the segmentation results, we also provide the affinity maps S^+ and S^- to illustrate the reasons of our improvement. It can be observed that, 1) The segmentation result is closely guided by the affinity map. Particularly, S^+ and S^- roughly describe the FG and BG objects, then segmentation algorithms refine the prediction based on the prior description. 2) GF-SAM + GSM achieves huge gain, because we greatly improve the affinity map. For example, in row 1, GF-SAM misses one of bird wing, and in row 2, it misses part of the bird tail. This is because its S^+ omits such regions guidance, and S^- also fails to clearly indicate the BG semantics. While our S^+ clearly indicates the whole wings of row 1 and whole tail of row 2, and S^- also clearly reveals the BG semantics, thus bringing better results. 3) Our affinity map improvement comes from two aspects, *i.e.,* clearer semantic representation and fewer outliers. For semantic representation, our approach achieves precise target localization and enhanced semantic discriminability. E.g. we have clearly higher and more continuous heat values for both FG and BG semantics in S^+ and S^-, respectively, which reveals the query's generalizing semantic can better indicate the complete target object. For outliers, *e.g.,* in row 3 and 4, the S^+ of GF-SAM have many low affinity points inside the FG region, while our outlier points are very few. This is precisely because RSM emphasizes the spatially continuous and region-wise distributed semantics.

Conclusion

In this paper, we propose GSM that focuses on exploiting generalizing semantic to improve the affinity calculation. We first organize the affinity-based inference into three main steps to reveal the crucial role of affinity map. Then, to address the low-data problem, TSR module considers the query sample as a proxy reference and assigns it with proxy mask identifying its most generalizing semantic regions. To generate high-fidelity proxy mask, QSM module pinpoints the most generalizing regions through prior semantic analysis. Finally, RR explicitly modulates affinity calculation via representation-aware weighting. Experiments on FSS benchmarks show that our GSM can serve as a plug-and-play free lunch for both specialist and foundation models.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 62176246. This work is also supported by Anhui Province Key Research and Development Plan (202304a05020045) and Anhui Province Natural Science Foundation (2208085UD17). This work is also supported by National Natural Science Foundation of China under Grant 62406098 and 62376256, and The Joint Fund for Medical Artificial Intelligence under Grant MAI2022Q011. This work is also supported by the Graduate School Special Funding Program of the University of Science and Technology of China.

References

He, W. 2024. Apseg: Auto-prompt network for cross-domain few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23762–23772.

Herzog, J. 2024. Adapt before comparison: A new perspective on cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23605–23615.

Hong, S.; Cho, S.; Nam, J.; Lin, S.; and Kim, S. 2022. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, 108–126. Springer.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026.

Li, F.; Jiang, Q.; Zhang, H.; Ren, T.; Liu, S.; Zou, X.; Xu, H.; Li, H.; Yang, J.; Li, C.; et al. 2024. Visual in-context prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12861–12871.

Li, G.; Jampani, V.; Sevilla-Lara, L.; Sun, D.; Kim, J.; and Kim, J. 2021. Adaptive Prototype Learning and Allocation for Few-Shot Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8334–8343.

Li, M.; et al. 2025. Improving Zero-Shot Generalization for CLIP with Prompt Ensemble self-Distillation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Liu, Y. 2024. Matcher: Segment Anything with One Shot Using All-Purpose Feature Matching. In *ICLR*.

Min, J. 2021. Hypercorrelation Squeeze for Few-Shot Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6941–6952.

Ng, A.; et al. 2001. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14.

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

Park, S.; Lee, S.; Hyun, S.; Seong, H. S.; and Heo, J.-P. 2024. Task-disruptive background suppression for few-shot segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 4442–4449.

Park, S.; Lee, S.; Seong, H. S.; Yoo, J.; and Heo, J.-P. 2025. Foreground-covering prototype generation and matching for sam-aided few-shot segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6425–6433.

Peng, B. 2023. Hierarchical Dense Correlation Distillation for Few-Shot Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23641–23651.

Shi, X.; Wei, D.; Zhang, Y.; Lu, D.; Ning, M.; Chen, J.; Ma, K.; and Zheng, Y. 2022a. Dense Cross-Query-and-Support Attention Weighted Mask Aggregation for Few-Shot Segmentation. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 151–168. Cham: Springer Nature Switzerland. ISBN 978-3-031-20044-1.

Shi, X.; Wei, D.; Zhang, Y.; Lu, D.; Ning, M.; Chen, J.; Ma, K.; and Zheng, Y. 2022b. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *European Conference on Computer Vision*, 151–168. Springer.

Su, J. 2024. Domain-rectifying adapter for cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 24036–24045.

Wang, J. 2024a. Adaptive FSS: a novel few-shot segmentation framework via prototype enhancement. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 5463–5471.

Wang, J. 2024b. Rethinking prior information generation with clip for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3941–3951.

Wang, Y. 2023. Focus on Query: Adversarial Mining Transformer for Few-Shot Segmentation. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 31524–31542. Curran Associates, Inc.

Wang, Y.; Sun, R.; and Zhang, T. 2023. Rethinking the Correlation in Few-Shot Segmentation: A Buoys View. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7183–7192.

Xiao, K.; et al. 2024. Semantic-guided robustness tuning for few-shot transfer across extreme domain shift. In *European Conference on Computer Vision*, 303–320. Springer.

Xu, Q.; Lin, G.; Loy, C. C.; Long, C.; Li, Z.; and Zhao, R. 2024. Eliminating feature ambiguity for few-shot segmentation. In *European Conference on Computer Vision*, 416–433. Springer.

Zhang, A. 2024a. Bridge the points: Graph-based few-shot segment anything semantically. *Advances in Neural Information Processing Systems*, 37: 33232–33261.

Zhang, R. 2024b. Personalize Segment Anything Model with One Shot. In *The Twelfth International Conference on Learning Representations*.

Zhu, L.; Chen, T.; Yin, J.; See, S.; and Liu, J. 2024. Addressing background context bias in few-shot segmentation through iterative modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3370–3379.