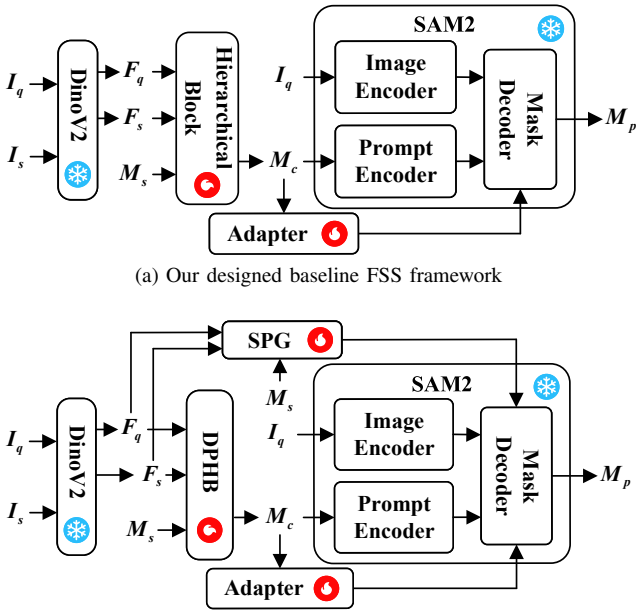
EasyFSS: Dual-Polarity Matching and Multi-Prompt Guidance for Few-Shot Segmentation

Yang Shi, Linsen Yu, Guanglu Sun*
College of Computer Science and Technology, Harbin University Of Science And Technology, Harbin 150080, China
2420410147@stu.hrbust.edu.cn, yulinsen@hrbust.edu.cn, sunguanglu@hrbust.edu.cn

***Abstract*—Few-Shot Segmentation (FSS) aims to recognize and segment novel classes in images using only a limited number of annotated examples. Recently, increasing efforts have explored incorporating foundation models (e.g., DINOv2 and SAM) into FSS. However, existing pixel-wise matching approaches based on DINOv2 features often rely solely on foreground-constrained modeling, resulting in less discriminative similarity responses and suboptimal dense prompts. Moreover, relying solely on dense prompts remains inadequate for effectively guiding the Segment Anything Model (SAM) family in constraining object boundaries and low-confidence regions. To address these challenges, we propose EasyFSS, a SAM2-based few-shot segmentation framework that explicitly enhances dense prompts quality while introducing complementary sparse prompts. Specifically, based on multi-level DINOv2 features, we construct a Dual-Polarity Hierarchical Block (DPHB), extending foreground-constrained modeling to foreground–background polarity matching across semantic hierarchies. This design significantly enhances the discriminability of pixel-wise similarity responses and alleviates background-induced ambiguity. Furthermore, we design a Sparse Prompt Generation (SPG) module that leverages global prototypes and multi-scale spatial context to generate high-confidence sparse prompts, effectively constraining boundaries and low-confidence regions. Finally, dense and sparse prompts collaboratively guide the SAM2 decoding process, enabling accurate and reliable segmentation predictions. Extensive qualitative and quantitative experiments demonstrate that EasyFSS achieves state-of-the-art performance, with 1-shot mIoU improvements of 3.2% and 3.9% on PASCAL-5^i and COCO-20^i, respectively. Our code is publicly available at https://github.com/stoneoceam/EasyFSS.**

***Index Terms*—Few-Shot Segmentation, SAM2, Dual-Polarity Matching, Multi-Prompt Guidance.**



(a) Our designed baseline FSS framework

(b) Our proposed EasyFSS framework

Fig. 1. Comparison between the baseline and our EasyFSS. (a) The baseline framework, based on SAM2 and inspired by prior works [7], [8]. (b) Our EasyFSS enhances the baseline by incorporating DPHB and SPG modules.

I. Introduction

Image segmentation is a fundamental task in computer vision, aiming to assign a semantic label to each pixel in an image. However, conventional segmentation methods [1]–[3] typically rely on large-scale, finely annotated datasets to achieve satisfactory performance, and they often struggle to generalize to unseen classes, which significantly limits their practical applicability. Inspired by the human ability to recognize novel objects from only a few examples, few-shot segmentation (FSS) [4]–[6] has emerged as a promising paradigm. FSS aims to guide models using only a small set of annotated support samples, enabling rapid generalization to segment novel classes in query images.

Early FSS methods can be categorized into prototype-based and affinity-based approaches. Prototype-based methods [5], [6], [9]–[14] extract representative vectors from support features via global average pooling. While they effectively suppress background noise, fine-grained structural details may be lost. Affinity-based methods [15]–[17] aim to establish dense pixel-level correspondences and may offer greater potential in handling complex backgrounds or significant intra-class variations. Although these methods have achieved notable progress under data-limited settings, their feature representation capacity is typically constrained by the employed pre-trained backbone networks. In complex scenarios with a large number of classes, the extracted features may not offer sufficient discriminative power, potentially constraining the model's segmentation performance.

In recent years, large-scale self-supervised pretrained foundation models, represented by DINOv2 [18], have demonstrated remarkable semantic representation capabilities, while

* Corresponding author

This research is partially supported by the Heilongjiang Provincial Discipline Innovation Project (Grant No. LJGXCG2024-F10), the National Science and Technology Major Project (Grant No. 2025ZD1607200), and the National Natural Science Foundation of China (Grant No. 62541604).

the emergence of the Segment Anything Model (SAM) family [19], [20] has significantly improved segmentation quality and generalization. Inspired by FounFSS [7] and CMaP-SAM [8], we design a baseline framework, as illustrated in Fig. 1(a). Specifically, we adopt the coarse masks produced by the hierarchical block in FounFSS, and align them into the dense prompts embedding space following the adapter-based strategy in CMaP-SAM. The aligned dense prompts are then used to drive SAM2 for final mask generation. However, this baseline still suffers from certain limitations. On the one hand, the matching strategy in FounFSS [7] primarily relies on support foreground features, without explicitly modeling potentially discriminative information from the support background. In scenes with complex backgrounds, such a design tends to weaken the spatial discriminability of the generated coarse masks, making them more prone to spurious foreground responses. Furthermore, when the coarse masks are further used to construct dense prompts, these noisy responses may propagate to the SAM2 decoding stage, thereby affecting its ability to perform fine-grained segmentation of target objects. On the other hand, although the SAM family [19], [20] exhibits strong decoding capability, its segmentation performance remains highly dependent on the quality of prompts. Relying solely on densely matched prompts is insufficient to effectively constrain low-confidence predictions along object boundaries and transition regions, which leads to suboptimal segmentation results.

To address these challenges, we present EasyFSS, a novel few-shot segmentation model that leverages DINOv2 and SAM2, as illustrated in Fig. 1(b). Specifically, we construct the Dual-Polarity Hierarchical Block (DPHB). DPHB explicitly incorporates background matching constraints and leverages both support foreground and background information to enhance the spatial discriminability of similarity responses, effectively reducing false-positive predictions in the coarse masks, thereby improving the reliability of the dense prompts. Although DPHB effectively improves pixel-wise matching, it still lacks explicit global prior constraints. To address this, we design the Sparse Prompt Generation (SPG) module. SPG generates a global prior using a prototype-based approach, models spatial context via different-scale features, and encodes it as sparse prompts. These prompts further guide SAM2 to focus on regions in the DPHB output that exhibit higher semantic consistency and stronger spatial confidence, enabling the mask decoder of SAM2 to produce more precise and fine-grained segmentation results.

The main contributions of this work are summarized as follows:

- We propose DPHB, which explicitly incorporates foreground and background polarity matching constraints during the correlation modeling stage, enhancing the spatial discriminability of pixel-wise similarity responses and producing more reliable coarse masks and dense prompts.
- We design a SPG module that leverages global prototypes and multi-scale spatial context to generate high-confidence sparse prompts, providing complementary global guidance to SAM2 and further improving segmentation performance.
- Extensive experimental results demonstrate that the proposed EasyFSS achieves state-of-the-art (SOTA) performance on both PASCAL-5^i and COCO-20^i benchmarks.

II. Related Works

A. Classic Few-Shot Segmentation

Few-shot Segmentation (FSS) extends the paradigm of few-shot learning to the semantic segmentation task, aiming to accurately segment novel classes given only a few annotated samples. Early studies predominantly employed prototype-based methods, which aggregate support features into a single prototype vector [5], [6], [9] and classify query pixels based on these prototypes. While such methods are simple and exhibit some generalization capability, compressing support information into a single vector may lead to incomplete or biased class representations. To alleviate this limitation, several approaches have been proposed, including multi-prototype strategies [10], [11], prototype enhancement methods [12], [13], and frequency-based prototype expansion techniques [14], all aiming to overcome the shortcomings of traditional prototype-based approaches in complex scenarios. Affinity-based methods [15]–[17] directly compute similarity between support and query features at the feature map level to construct dense correlation maps. These methods are more effective at capturing fine-grained spatial correspondences, enabling more precise localization of target regions. Moreover, some approaches [21], [22] attempt to integrate prototype-based and affinity-based methods, leveraging their complementary strengths to further improve segmentation performance. OCNet [23] introduces object-level correlations to address the limitations of image-level correlations. Although these methods improve segmentation performance to some extent, their effectiveness is still constrained by the limited semantic representation capacity of pre-trained backbone networks [24], [25], leading to performance bottlenecks when handling extremely complex scenes.

B. Foundation-based Few-Shot Segmentation

With the advent of large-scale foundation models, the research paradigm of FSS has undergone a significant transformation. Self-supervised Vision Transformer models, such as DINOv2 [18], possess powerful semantic feature representation capabilities and exhibit clear advantages in cross-category generalization compared to traditional CNN backbones [24], [25]. Meanwhile, an increasing number of studies have explored integrating prompt-driven SAM models [19], [20] into the FSS task. GF-SAM [26] generates point prompts via automatic clustering to drive SAM for segmentation. CMaP-SAM [8] introduces an adaptive distribution alignment module that aligns initial segmentation masks to the prompt embedding space of SAM, thereby enhancing the representational power of dense prompts. FCP [27] constructs separate support and query prototypes and compares them to generate visual reference prompts that guide SAM. FSSAM [28] further

leverages SAM2's memory mechanism and employs a pseudo-prompt generator to mitigate performance degradation caused by intra-class appearance variations. Although these methods improve the adaptability of the SAM family to few-shot segmentation to varying degrees, most still rely on single or coarse-grained prompts and insufficiently model discriminative information between foreground and background, making them prone to mis-segmentation in complex scenarios. To address these limitations, we refine the dense and sparse prompt generation strategies to further exploit the complementarity between semantic features and prompt information, thereby enhancing the few-shot segmentation capability of the unified DINOv2 and SAM2 framework.

III. Methodology

A. Problem Definition

Few-Shot Segmentation (FSS) aims to perform pixel-level segmentation for unseen categories given only a limited number of annotated samples. In this setting, the training set $\boldsymbol{D}_{\text{train}}$ and test set $\boldsymbol{D}_{\text{test}}$ correspond to two disjoint category sets, $\boldsymbol{C}_{\text{base}}$ and $\boldsymbol{C}_{\text{novel}}$, with $\boldsymbol{C}_{\text{base}} \cap \boldsymbol{C}_{\text{novel}} = \varnothing$, ensuring that test categories are never seen during training. To address this challenge, FSS typically adopts an episodic learning paradigm, where each episode consists of a support set $\boldsymbol{S} = (\boldsymbol{I}_s^i, \boldsymbol{M}_s^i)_{i=1}^k$, containing k support samples for a target category, and a query sample $(\boldsymbol{I}_q, \boldsymbol{M}_q)$. Under this framework, the model exploits the limited prior information in the support set to segment the query image and learns transferable segmentation capabilities over $\boldsymbol{C}_{\text{base}}$, enabling generalization to unseen categories in $\boldsymbol{C}_{\text{novel}}$ at test time without additional parameter updates.

B. Overview

An overview of our model EasyFSS is illustrated in Fig. 2. First, DINOv2 is used to extract features from both the support and query images, yielding support features $\boldsymbol{F}_s$ and query features $\boldsymbol{F}_q$, respectively. Based on these features, the Dual-Polarity Hierarchical Block (DPHB) jointly models foreground and background semantic correlations between $\boldsymbol{F}_q$ and $\boldsymbol{F}_s$, generating coarse masks $\boldsymbol{M}_c$ as a structural prior. This prior is further processed by an adapter to produce adapted dense prompt features, which are combined with SAM2 dense prompt embeddings to form dense prompts. In addition, the Sparse Prompt Generator (SPG) module takes the top-level features $\boldsymbol{F}_q^{top}$ and $\boldsymbol{F}_s^{top}$ from the query and support features as input, extracts globally discriminative semantic cues, and generates sparse prompts to capture key object locations and discriminative information. Finally, we discard the memory-related modules of SAM2 and inject the dense prompts from DPHB and the sparse prompts from SPG into its mask decoding process. During decoding, these prompts collaboratively guide SAM2 to refine the coarse segmentation, resulting in high-quality final segmentation predictions.

C. Feature Extraction

Following [7], we employ the pre-trained DINOv2 encoder $\mathcal{E}_{\text{DINOv2}}(\cdot)$ to extract hierarchical semantic features from the support image $\boldsymbol{I}_s$ and the query image $\boldsymbol{I}_q$. The outputs of the l-th self-attention layer are denoted as $\boldsymbol{F}_s^l$ and $\boldsymbol{F}_q^l$, respectively. This process can be expressed as:

$$\left\{\boldsymbol{F}_q^l\right\}_{l=1}^{L} = \mathcal{E}_{\text{DINOv2}}(\boldsymbol{I}_q), \quad \left\{\boldsymbol{F}_s^l\right\}_{l=1}^{L} = \mathcal{E}_{\text{DINOv2}}(\boldsymbol{I}_s), \tag{1}$$

where $\boldsymbol{I}_q, \boldsymbol{I}_s \in \mathbb{R}^{3\times H\times W}$, $l \in \{1, \ldots, L\}$ with $L = 12$, and $\boldsymbol{F}_q^l, \boldsymbol{F}_s^l \in \mathbb{R}^{C\times h\times w}$ denote the reshaped patch-level feature maps at the corresponding spatial resolution.

D. Dual-Polarity Hierarchical Block

Existing matching-based methods [7], [15], [29] primarily focus on the similarity between query features and support foregrounds, implicitly assuming that regions with high similarity correspond to the target. However, in scenarios with complex backgrounds or large intra-class variations, relying solely on foreground similarity is insufficient to provide stable discriminative cues at the spatial level, often leading to false-positive responses and resulting in segmentation errors. In contrast, the support background contains important discriminative information about non-target appearances, which is typically not explicitly modeled in existing correlation frameworks and only serves as an implicit constraint. To address this limitation, we propose the Dual-Polarity Hierarchical Block (DPHB), which explicitly incorporates both foreground and background matching constraints during correlation modeling. This enhances the spatial discriminability of pixel-wise similarity responses, providing more reliable guidance for subsequent segmentation.

Specifically, as illustrated in Fig. 2, for the l-th feature layer, we first extract the support foreground features $\boldsymbol{F}_s^{fg,l}$ and support background features $\boldsymbol{F}_s^{bg,l}$ using the support mask $\boldsymbol{M}_s$. This process is formulated as:

$$\boldsymbol{F}_s^{fg,l} = \boldsymbol{F}_s^l \odot \boldsymbol{M}_s, \quad \boldsymbol{F}_s^{bg,l} = \boldsymbol{F}_s^l \odot (1 - \boldsymbol{M}_s), \tag{2}$$

where $\odot$ denotes the Hadamard product (element-wise multiplication).

Subsequently, in each feature layer, we measure the cosine similarity between the query features $\boldsymbol{F}_q^l$ and the support foreground features $\boldsymbol{F}_s^{fg,l}$, as well as between the query features $\boldsymbol{F}_q^l$ and the support background features $\boldsymbol{F}_s^{bg,l}$, resulting in the following correlation maps:

$$\boldsymbol{H}_{q,s}^{fg,l} = \text{Cos}(\boldsymbol{F}_q^l, \boldsymbol{F}_s^{fg,l}), \quad \boldsymbol{H}_{q,s}^{bg,l} = \text{Cos}(\boldsymbol{F}_q^l, \boldsymbol{F}_s^{bg,l}). \tag{3}$$

Here, we denote the resulting 4D correlation space of size $h \times w \times h \times w$ as $\boldsymbol{\Omega}$. $\boldsymbol{H}_{q,s}^{fg,l} \in \mathbb{R}^{h\times w\times h\times w}$ and $\boldsymbol{H}_{q,s}^{bg,l} \in \mathbb{R}^{h\times w\times h\times w}$ represent the correlations between the l-th layer query features and the foreground and background regions of the support features, respectively. To capture hierarchical semantic information, we aggregate the per-layer correlation maps into foreground and background correlation sets across all L semantic levels, which are defined as:

$$\boldsymbol{H}_{q,s}^{fg} = \left\{\boldsymbol{H}_{q,s}^{fg,l}\right\}_{l=1}^{L}, \quad \boldsymbol{H}_{q,s}^{fg} \in \mathbb{R}^{L\times\boldsymbol{\Omega}}, \tag{4}$$

$$\boldsymbol{H}_{q,s}^{bg} = \left\{\boldsymbol{H}_{q,s}^{bg,l}\right\}_{l=1}^{L}, \quad \boldsymbol{H}_{q,s}^{bg} \in \mathbb{R}^{L\times\boldsymbol{\Omega}}, \tag{5}$$

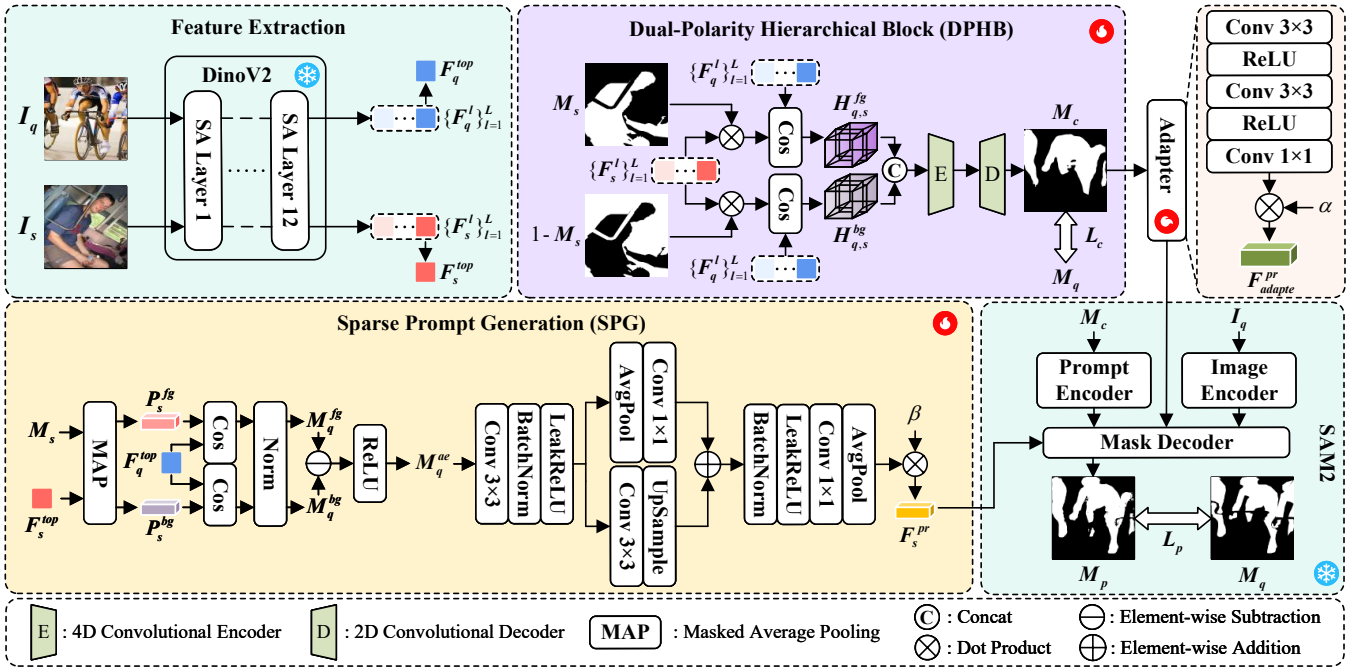


Fig. 2. Overview of our method. The Dual-Polarity Hierarchical Block (DPHB) constructs bipolar semantic correlation representations from multi-level DINOv2 features and generates coarse masks as a structural prior, which is then mapped via an adapter into dense semantic prompts to provide continuous spatial constraints. The Sparse Prompt Generation (SPG) module extracts highly discriminative semantic prototypes from top-level DINOv2 features to form global constraints and produce sparse semantic prompts. Finally, the dense and sparse semantic prompts jointly guide the decoding process of SAM2, enhancing both the accuracy and stability of segmentation predictions.

where the foreground correlation set $\boldsymbol{H}_{q,s}^{fg}$ characterizes the similarity between query regions and the target appearance across multiple semantic levels, whereas the background correlation set $\boldsymbol{H}_{q,s}^{bg}$ captures matching relationships with non-target appearances. We concatenate the hierarchical foreground and background correlation sets along the semantic level dimension to form the bipolar correlation representation $\boldsymbol{H}_{q,s}$. This process can be expressed as:

$$\boldsymbol{H}_{q,s} = \mathrm{Concat}\left(\boldsymbol{H}_{q,s}^{fg}, \boldsymbol{H}_{q,s}^{bg}\right), \quad \boldsymbol{H}_{q,s} \in \mathbb{R}^{2L\times\boldsymbol{\Omega}}. \tag{6}$$

Finally, following [7], we encode and decode $\boldsymbol{H}_{q,s}$ to obtain the coarse masks $\boldsymbol{M}_c$, as follows:

$$\boldsymbol{M}_c = \mathrm{D}(\mathrm{E}(\boldsymbol{H}_{q,s})), \tag{7}$$

where $\mathrm{E}(\cdot)$ denotes a 4D convolutional encoder, $\mathrm{D}(\cdot)$ denotes a 2D convolutional decoder, and $\boldsymbol{M}_c \in \mathbb{R}^{1\times H\times W}$.

E. Sparse Prompt Generation

Although DPHB significantly enhances the discriminability of similarity responses during the correlation modeling stage and produces coarse masks $\boldsymbol{M}_c$ with relatively high coverage, these masks are still mainly driven by dense matching responses and lack constraints from global priors. When only dense prompts are used to drive SAM2, the resulting segmentation outputs often exhibit low-confidence responses along object boundaries and transitional regions. Inspired by [21], we further design a Sparse Prompt Generation (SPG) module, which generates priors with global discriminability using a prototype-based approach and encodes them as sparse prompts by modeling multi-scale spatial context, thereby guiding SAM2 to focus on regions in $\boldsymbol{M}_c$ with higher semantic consistency and spatial confidence.

As shown in Fig. 2, we construct a prior query mask $\boldsymbol{M}_q^{ae}$ in the SPG module following [30]. Specifically, given the top-level query features $\boldsymbol{F}_q^{top}$, top-level support feature $\boldsymbol{F}_s^{top}$, and support mask $\boldsymbol{M}_s$, we perform Masked Average Pooling (MAP) over the top-level support features. This aggregates the foreground and background feature representations in the support, yielding the support foreground prototype $\boldsymbol{P}_s^{fg}$ and support background prototype $\boldsymbol{P}_s^{bg}$. These processes are represented as:

$$\boldsymbol{P}_s^{fg} = \mathrm{MAP}(\boldsymbol{F}_s^{top}, \boldsymbol{M}_s), \tag{8}$$

$$\boldsymbol{P}_s^{bg} = \mathrm{MAP}(\boldsymbol{F}_s^{top}, 1-\boldsymbol{M}_s). \tag{9}$$

Next, the contrastive responses between the top-level query feature and the foreground/background prototypes are computed via cosine similarity, followed by normalization to obtain the prior query mask $\boldsymbol{M}_q^{ae}$. The prior query mask is expressed as follows:

$$\boldsymbol{M}_q^{ae} = \varphi\left(\mathcal{N}\left(\mathrm{Cos}(\boldsymbol{F}_q^{top}, \boldsymbol{P}_s^{fg})\right) - \mathcal{N}\left(\mathrm{Cos}(\boldsymbol{F}_q^{top}, \boldsymbol{P}_s^{bg})\right)\right), \tag{10}$$

where $\mathcal{N}(\cdot)$ denotes the min–max scaler to normalize the values into $[0, 1]$, $\varphi(\cdot)$ denotes the ReLU activation function, and $\boldsymbol{M}_q^{ae} \in \mathbb{R}^{1\times H\times W}$.

We then refine $\boldsymbol{M}_q^{ae}$ at the prompts feature level to enhance its representation of local contextual information and spatial structural patterns. Specifically, we first extract spatial dependencies among neighboring locations in $\boldsymbol{M}_q^{ae}$, producing a spatially refined prompts representation $\boldsymbol{P}_q^{ref}$. This step is expressed as:

$$\boldsymbol{P}_q^{ref} = \phi_{\text{ref}}(\boldsymbol{M}_q^{ae}), \tag{11}$$

where $\phi_{\text{ref}}(\cdot)$ consists of a 3×3 convolution, Batch Normalization, and LeakyReLU.

To balance global semantic consistency and local structural discriminability, we further aggregate the refined prompts representation using a multi-scale context aggregation, yielding the aggregated feature $\boldsymbol{F}_q^{agg}$, which is defined as:

$$\boldsymbol{F}_q^{agg} = \phi_{\text{global}}(\boldsymbol{P}_q^{ref}) + \phi_{\text{local}}(\boldsymbol{P}_q^{ref}), \tag{12}$$

where $\phi_{\text{global}}(\cdot)$ represents the global modeling branch realized by average pooling followed by a 1×1 convolution, and $\phi_{\text{local}}(\cdot)$ represents the local structural modeling branch realized by a 3×3 convolution and an upsampling operation.

Finally, the aggregated feature is mapped into the SAM2 prompt space to produce the sparse prompts $\boldsymbol{F}_s^{pr}$. This process can be formally expressed as:

$$\boldsymbol{F}_s^{pr} = \tanh(\boldsymbol{\beta}) \cdot \phi_{\text{cal}}(\boldsymbol{F}_q^{agg}), \tag{13}$$

where $\phi_{\text{cal}}(\cdot)$ consists of Batch Normalization, LeakyReLU, a 1×1 convolution, and average pooling, with $\boldsymbol{\beta}$ being a learnable parameter initialized to 0 and bounded by $\tanh(\cdot)$.

F. Final prediction

We extract the SAM2 query image features from the query image $\boldsymbol{I}_q$ using the SAM2 image encoder, denoted as $\boldsymbol{F}_q^{\text{SAM2}}$, which is defined as:

$$\boldsymbol{F}_q^{\text{SAM2}} = \mathcal{E}_{\text{SAM2}}(\boldsymbol{I}_q), \tag{14}$$

where $\mathcal{E}_{\text{SAM2}}(\cdot)$ represents the image encoder of SAM2.

Following [8], the coarse masks $\boldsymbol{M}_c$ are processed by an adaptive alignment module $A(\cdot)$, yielding the adapted dense prompt features $\boldsymbol{F}_{adapt}^{pr}$. This process can be formulated as:

$$\boldsymbol{F}_{adapt}^{pr} = \boldsymbol{\alpha} \cdot A(\boldsymbol{M}_c), \tag{15}$$

where $\boldsymbol{\alpha}$ is a learnable scaling parameter initialized to 1, and $A(\cdot)$ consists of convolutional and ReLU layers.

Next, the dense prompts $\boldsymbol{F}_d^{pr}$ are obtained by combining these adapted features with the output of SAM2's Prompt Encoder $P_{\text{SAM2}}(\cdot)$. This step can be expressed as:

$$\boldsymbol{F}_d^{pr} = P_{\text{SAM2}}(\boldsymbol{M}_c) + \boldsymbol{F}_{adapt}^{pr}. \tag{16}$$

Finally, the SAM2 query image feature $\boldsymbol{F}_q^{\text{SAM2}}$, the dense prompts $\boldsymbol{F}_d^{pr}$, and the sparse prompts $\boldsymbol{F}_s^{pr}$ are fed into the SAM2 Mask Decoder to produce final predicted masks $\boldsymbol{M}_p$. The final predicted masks are obtained as:

$$\boldsymbol{M}_p = D_{\text{SAM2}}(\boldsymbol{F}_q^{\text{SAM2}}, \boldsymbol{F}_d^{pr}, \boldsymbol{F}_s^{pr}), \tag{17}$$

where $D_{\text{SAM2}}(\cdot)$ denotes the Mask Decoder of SAM2, and $\boldsymbol{M}_p \in \mathbb{R}^{1\times H\times W}$.

G. Loss Function

During training, supervision is applied to both the coarse masks $\boldsymbol{M}_c$ and the predicted masks $\boldsymbol{M}_p$. We adopt a combination of Dice Loss and Binary Cross Entropy (BCE) Loss to ensure both coarse localization and fine-grained segmentation accuracy. Specifically, the loss functions are defined as:

$$\boldsymbol{L}_c = L_{\text{Dice}}(\boldsymbol{M}_c, \boldsymbol{M}_q) + L_{\text{BCE}}(\boldsymbol{M}_c, \boldsymbol{M}_q), \tag{18}$$

$$\boldsymbol{L}_p = L_{\text{Dice}}(\boldsymbol{M}_p, \boldsymbol{M}_q) + L_{\text{BCE}}(\boldsymbol{M}_p, \boldsymbol{M}_q), \tag{19}$$

$$\boldsymbol{L} = \boldsymbol{L}_c + \boldsymbol{L}_p. \tag{20}$$

Here, $\boldsymbol{M}_q$ denotes the ground-truth query mask. $L_{\text{Dice}}(\cdot)$ represents the Dice Loss, which mitigates class imbalance and improves region overlap, while $L_{\text{BCE}}(\cdot)$ represents the BCE Loss, enforcing pixel-wise classification consistency. $\boldsymbol{L}$ denotes the final loss.

H. Extending to K-shot setting

Following [7], [15], for each query image, we independently perform segmentation using $k\,(k > 1)$ support samples to obtain k predicted masks, which are then averaged and thresholded with $\tau = 0.5$ to produce the final segmentation result.

IV. Experiments

A. Datasets

We evaluate the proposed method on two widely used benchmarks, PASCAL-5^i [4] and COCO-20^i [33]. PASCAL-5^i, built upon PASCAL VOC 2012 [34] with additional SDS annotations [35], contains 20 categories, while COCO-20^i is derived from MS COCO [36] with 80 categories and presents greater challenges due to its category diversity and complex backgrounds. Following the standard cross-validation protocol, both datasets are split into four disjoint folds, each containing 5 and 20 categories for PASCAL-5^i and COCO-20^i, respectively. Three folds are used for training, while the remaining fold is reserved for testing.

B. Implementation Details

We use DINOv2 ViT-B/16 as the feature extractor and adopt the SAM2-Hiera Base-plus model as the segmentation backbone. All experiments are conducted on an NVIDIA 4090 GPU. During training, the batch size is set to 20, and the AdamW optimizer [37] is employed with a learning rate of 1e-3 and a weight decay of 5e-4. The resolutions of both support and query images are resized to 420×420 pixels. Our model is trained for 30 epochs on the PASCAL-5^i dataset and for 15 epochs on the COCO-20^i dataset, respectively, with the final model selected based on validation performance. Notably, no additional data augmentation strategies are applied except for the resizing operation.

C. Evaluation Metrics

We utilize mean intersection over union (mIoU) and foreground-background IoU (FB-IoU) as evaluation metrics.

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON PASCAL-5^i IN TERMS OF MIOU AND FB-IOU. THE BEST AND SECOND-BEST PERFORMANCES ARE IN **BOLD** AND UNDERLINED, RESPECTIVELY.

Method	Publication	1-shot						5-shot					
		Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FB-IoU	Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FB-IoU
Classical FSS													
ABCB [31]	CVPR'25	73.0	76.0	69.7	69.2	72.0	–	74.8	78.5	73.6	72.6	74.9	–
PAHNet [21]	ICCV'25	72.1	76.4	71.0	66.6	71.5	82.4	75.8	78.6	76.9	73.2	76.1	85.7
OCNet [23]	ICCV'25	73.5	75.9	71.1	64.9	71.4	82.2	75.9	77.1	74.1	70.9	74.5	84.7
Foundation-based FSS													
Matcher [32]	ICLR'24	67.7	70.7	66.9	67.0	68.1	–	71.4	77.5	74.1	72.8	74.0	–
GF-SAM [26]	NeurIPS'24	71.1	75.7	69.2	73.3	72.1	–	81.5	86.3	79.7	82.9	82.6	–
FounFSS [7]	arXiv'24	76.5	81.3	72.1	77.4	76.8	–	79.5	84.8	75.8	82.5	80.7	–
FCP [27]	AAAI'25	74.9	77.4	71.8	68.8	73.2	–	77.2	78.8	72.2	67.7	74.0	–
CMaP-SAM [8]	Neurocomputing'25	78.1	75.5	65.0	66.0	71.1	80.8	82.2	81.2	68.5	73.2	76.5	85.1
FSSAM [28]	ICML'25	81.6	84.9	81.6	76.0	81.0	89.4	84.1	88.5	**83.8**	85.0	85.4	91.9
Ours	–	**82.6**	**85.6**	**82.7**	**86.0**	**84.2**	**90.8**	**84.1**	**88.5**	83.0	**88.1**	**85.9**	**92.0**

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON COCO-20^i IN TERMS OF MIOU AND FB-IOU. THE BEST AND SECOND-BEST PERFORMANCES ARE IN **BOLD** AND UNDERLINED, RESPECTIVELY.

Method	Publication	1-shot						5-shot					
		Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FB-IoU	Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FB-IoU
Classical FSS													
ABCB [31]	CVPR'25	46.0	56.3	54.3	51.3	51.5	–	51.6	63.5	62.8	57.2	58.8	–
PAHNet [21]	ICCV'25	45.2	58.2	54.4	51.2	52.3	75.2	53.0	64.1	60.8	58.7	59.2	78.1
OCNet [23]	ICCV'25	45.9	56.9	52.9	50.4	51.5	73.7	52.7	63.1	57.4	54.8	57.0	76.8
Foundation-based FSS													
Matcher [32]	ICLR'24	52.7	53.5	52.6	52.1	52.7	–	60.1	62.7	60.9	59.2	60.7	–
GF-SAM [26]	NeurIPS'24	56.6	61.4	59.6	57.1	58.7	–	67.1	69.4	66.0	64.8	66.8	–
FounFSS [7]	arXiv'24	56.0	61.3	57.9	58.8	58.5	–	61.4	69.4	65.9	64.9	65.4	–
FCP [27]	AAAI'25	46.4	56.4	55.3	51.8	52.5	–	52.6	63.3	59.8	56.1	58.0	–
CMaP-SAM [8]	Neurocomputing'25	54.0	56.1	58.9	55.3	56.1	75.8	61.3	66.8	67.3	65.9	65.3	81.1
FSSAM [28]	ICML'25	59.9	65.6	62.1	61.6	62.3	77.3	**68.6**	**74.0**	64.5	69.9	69.3	82.9
Ours	–	**64.6**	**68.0**	**66.6**	**65.7**	**66.2**	**81.4**	67.3	73.2	**72.3**	**70.6**	**70.8**	**83.9**

D. Comparisons with State-of-the-Arts

To evaluate the effectiveness of the proposed method, we conduct comparative experiments on two benchmark datasets, PASCAL-5^i and COCO-20^i, using mIoU and FB-IoU as evaluation metrics. The results are summarized in Table I and Table II. Specifically, on the PASCAL-5^i dataset, our method achieves an mIoU of 84.2% and an FB-IoU of 90.8% in the 1-shot setting, which are 3.2% and 1.4% higher than those of the previous best-performing method FSSAM [28], respectively. In the 5-shot setting, our method attains an mIoU of 85.9% and an FB-IoU of 92.0%, achieving the best performance across most folds and demonstrating the robustness of our model under varying class distributions. On the more challenging and diverse COCO-20^i dataset, our method achieves an mIoU of 66.2% and an FB-IoU of 81.4% in the 1-shot setting, and 70.8% and 83.9% in the 5-shot setting, which are 3.9%, 4.1% and 1.5%, 1.0% higher than FSSAM [28], respectively. These results demonstrate that our model, EasyFSS, achieves robust and precise segmentation even under limited-sample and diverse semantic conditions.

E. Qualitative Results

As shown in Fig. 3, we present qualitative segmentation results to further demonstrate the effectiveness of the proposed modules in challenging few-shot scenarios. The baseline method performs poorly when objects exhibit large structural variations or thin and fragmented shapes. Since it mainly relies on foreground similarity matching without explicitly modeling foreground–background relationships, the predicted masks contain scattered responses and incomplete structures. Specifically, for "Kite" and "Vase", the baseline produces fragmented predictions and fails to capture complete object shapes. For "Potted Plant", strong background interference causes excessive activations around leaves and branches, leading to noisy masks. For "Chair", the predicted regions lack structural consistency and the legs are poorly separated. After introducing the DPHB module, dual-polarity foreground–background constraints suppress irrelevant responses and improve tar-

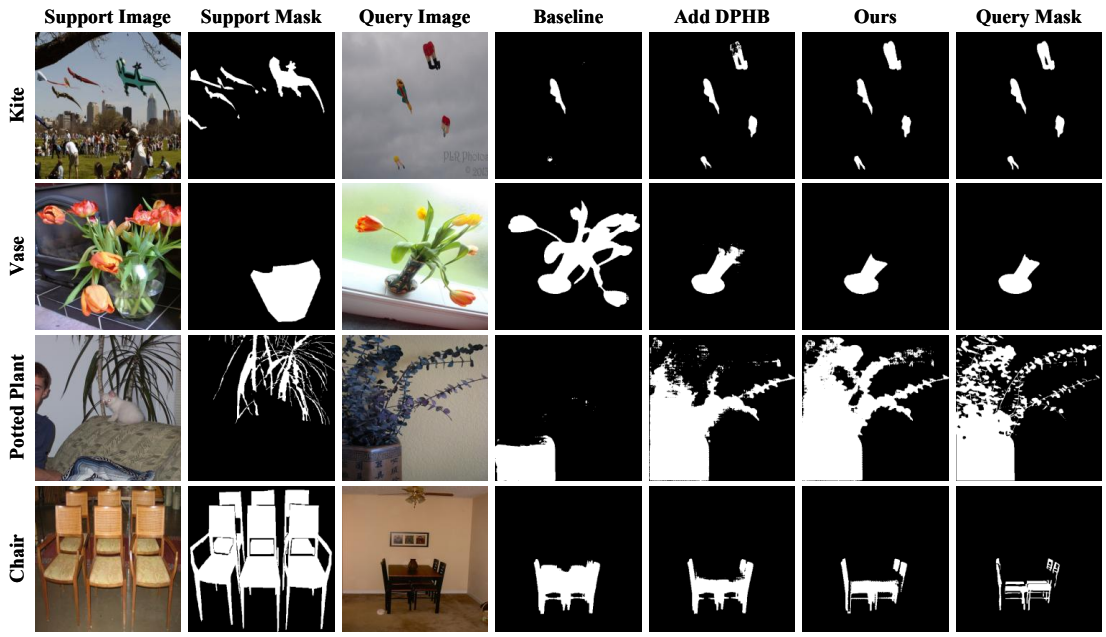


Fig. 3. Qualitative analysis results on representative categories. From left to right: query image, ground-truth mask, Baseline, Baseline + DPHB, and the complete model (Ours). Baseline often produces false positives in complex backgrounds, with mask holes or boundary artifacts on fine-grained objects. Introducing the DPHB module leverages dual foreground–background matching to effectively suppress background interference. Upon further incorporation of SPG module, the final model achieves significant improvements in edge fitting and mask consistency, producing masks that are closer to the ground truth.

TABLE III
COMPONENT-WISE ABLATION STUDY. THE RESULTS OF THE COMPLETE MODEL ARE HIGHLIGHTED IN **BOLD**.

Baseline	DPHB	SPG	F-0	F-1	F-2	F-3	mIoU	FB-IoU
✓			62.6	65.4	64.6	62.7	63.8	79.9
✓	✓		63.8	67.8	66.0	64.8	65.6	80.9
✓		✓	62.8	66.2	65.1	63.5	64.4	80.3
✓	✓	✓	**64.6**	**68.0**	**66.6**	**65.7**	**66.2**	**81.4**

get localization. Fragmented regions in "Kite" become more concentrated, "Vase" masks better capture the main body, background activations in "Potted Plant" are reduced, and the contours of "Chair" become clearer. By further integrating SPG, high-confidence sparse prompts guided by global context are generated to assist SAM2 in refining structures and boundaries. Consequently, "Kite" instances become more complete, "Vase" contours are smoother, leaf structures in "Potted Plant" are better preserved, and the "Chair" layout becomes more accurate with clearer separation between chair legs.

F. Ablation Study and Analysis

To further validate the effectiveness of the proposed module, we conducted an ablation study on the COCO-20^i dataset under the 1-shot setting.

1) Ablation Study on DPHB: The DPHB module explicitly introduces dual matching constraints between foreground and background during the correlation modeling stage, effectively enhancing the discriminability and stability of similarity responses in the spatial domain, thereby providing more reliable coarse masks for subsequent segmentation. As shown in Table III, compared with the baseline that relies solely on foreground similarity, incorporating background similarity leads to consistent performance improvements across all four folds, where the mIoU increases from 63.8% to 65.6% and the FB-IoU improves from 79.9% to 80.9%. These results indicate that explicitly modeling background information from the support set and jointly enforcing foreground–background correlation constraints plays an important role in suppressing false positive responses under complex backgrounds, and is a key factor in constructing robust matching relationships.

2) Ablation Study on SPG: Although DPHB is already capable of effectively driving SAM2 and establishing a strong segmentation baseline, the coarse masks M_c it produces are still primarily driven by dense matching responses, which inevitably contain a certain proportion of ambiguous regions along object boundaries or semantic transition areas. To further provide SAM2 with global and high-confidence structural constraint prompts, we introduce the Sparse Prompt Generation (SPG) module. As shown in Table III, incorporating SPG on top of DPHB leads to consistent performance improvements across all four folds, with overall mIoU and FB-IoU increased by approximately 0.6% and 0.5%, respectively. These results clearly demonstrate that SPG can effectively select prompt regions with stronger semantic consistency and higher spatial confidence, thereby enhancing the model's discriminative ability in complex scenes and fine-grained local regions. It is worth emphasizing that even when built upon the strong baseline established by DPHB, SPG still brings stable and substantial performance gains, which verifies its indispensable role in improving model generalization and segmentation stability.

V. Conclusion

In this work, we propose EasyFSS, a SAM2-based few-shot segmentation framework that leverages dual-polarity matching and multi-prompt guidance. The proposed DPHB explicitly models complementary foreground–background correlations to enhance the discriminability of dense matching responses, while the SPG module provides high-confidence, prototype-driven sparse prompts to constrain boundaries and low-confidence regions. By integrating dense and sparse prompts as complementary guidance signals, EasyFSS enables SAM2 to achieve both accurate object localization and fine-grained mask refinement. Extensive experiments on PASCAL-5^i and COCO-20^i validate the effectiveness of dual-polarity matching and multi-prompt guidance for SAM2-based few-shot segmentation. Future work will explore incorporating textual information to further enhance segmentation performance under few-shot settings.

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[3] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[4] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," *arXiv preprint arXiv:1709.03410*, 2017.

[5] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9197–9206.

[6] 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 analysis and machine intelligence*, vol. 44, no. 2, pp. 1050–1065, 2020.

[7] S. Chang, L. Zhang, and H. Lu, "High-performance few-shot segmentation with foundation models: An empirical study," *arXiv preprint arXiv:2409.06305*, 2024.

[8] S. Chen, F. Meng, L. Lei, H. Wei, C. Wu, Q. Wu, L. Xu, and H. Li, "Cmap-sam: Contraction mapping prior for sam-driven few-shot segmentation," *Neurocomputing*, p. 131816, 2025.

[9] Q. Fan, W. Pei, Y.-W. Tai, and C.-K. Tang, "Self-support few-shot semantic segmentation," in *European conference on computer vision*. Springer, 2022, pp. 701–719.

[10] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, and J. Kim, "Adaptive prototype learning and allocation for few-shot segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8334–8343.

[11] R. Cong, H. Xiong, J. Chen, W. Zhang, Q. Huang, and Y. Zhao, "Query-guided prototype evolution network for few-shot segmentation," *IEEE Transactions on Multimedia*, vol. 26, pp. 6501–6512, 2024.

[12] J. Wang, J. Li, C. Chen, Y. Zhang, H. Shen, and T. Zhang, "Adaptive fss: a novel few-shot segmentation framework via prototype enhancement," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 6, 2024, pp. 5463–5471.

[13] H. Li, G. Huang, X. Yuan, Z. Zheng, X. Chen, G. Zhong, and C.-M. Pun, "Psanet: prototype-guided salient attention for few-shot segmentation," *The Visual Computer*, vol. 41, no. 4, pp. 2987–3001, 2025.

[14] C. Wen, H. Huang, Y. Ma, F. Yuan, and H. Zhu, "Dual-guided frequency prototype network for few-shot semantic segmentation," *IEEE Transactions on Multimedia*, vol. 26, pp. 8874–8888, 2024.

[15] J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6941–6952.

[16] A. Tavera, F. Cermelli, C. Masone, and B. Caputo, "Pixel-by-pixel cross-domain alignment for few-shot semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1626–1635.

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

[18] 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.

[19] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[20] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[21] T. Zou, S. Xiong, R. Yao, and Y. Rong, "Balancing conservatism and aggressiveness: Prototype-affinity hybrid network for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 20 561–20 571.

[22] L. Cao, Y. Guo, Y. Yuan, and Q. Jin, "Prototype as query for few shot semantic segmentation," *Complex & Intelligent Systems*, vol. 10, no. 5, pp. 7265–7278, 2024.

[23] C. Wen, Y. Zhang, J. Fan, H. Zhu, X.-S. Wei, Y. Wang, Z. Kou, and S. Sun, "Object-level correlation for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 23 689–23 699.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] A. Zhang, G. Gao, J. Jiao, C. Liu, and Y. Wei, "Bridge the points: Graph-based few-shot segment anything semantically," *Advances in Neural Information Processing Systems*, vol. 37, pp. 33 232–33 261, 2024.

[27] S. Park, S. Lee, H. S. Seong, J. Yoo, and J.-P. Heo, "Foreground-covering prototype generation and matching for sam-aided few-shot segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, 2025, pp. 6425–6433.

[28] Q. Xu, L. Zhu, X. Liu, G. Lin, C. Long, Z. Li, and R. Zhao, "Unlocking the power of sam 2 for few-shot segmentation," *arXiv preprint arXiv:2505.14100*, 2025.

[29] A. Karimi and C. Poullis, "Dsv-lfs: Unifying llm-driven semantic cues with visual features for robust few-shot segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 4584–4594.

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

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

[32] Y. Liu, M. Zhu, H. Li, H. Chen, X. Wang, and C. Shen, "Matcher: Segment anything with one shot using all-purpose feature matching," *arXiv preprint arXiv:2305.13310*, 2023.

[33] K. Nguyen and S. Todorovic, "Feature weighting and boosting for few-shot segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 622–631.

[34] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[35] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *European conference on computer vision*. Springer, 2014, pp. 297–312.

[36] 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 *European conference on computer vision*. Springer, 2014, pp. 740–755.

[37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.