Learning to Segment with Deep Models in Low-Data Regimes

Amin Karimi

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Computer science) at

Concordia University

Montréal, Québec, Canada

November 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mr. Amin Karimi**

Entitled: **Learning to Segment with Deep Models in Low-Data Regimes**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Computer science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Susan Liscouët-Hanke

_____ External Examiner
Dr. Guilherme N. DeSouza

_____ Examiner
Dr. Jamal Bentahar

_____ Examiner
Dr. Adam Krzyzak

_____ Supervisor
Dr. Charalambos Poullis

Approved by _____
Joey Paquet, Chair
Department of Computer Science and Software Engineering

_____ 2025

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Learning to Segment with Deep Models in Low-Data Regimes

Amin Karimi, Ph.D.

Concordia University, 2025

This thesis addresses the challenge of few-shot semantic segmentation (FSS), aiming to achieve accurate image understanding in low-data regimes. Traditional few-shot semantic segmentation methods often struggle to generalize effectively, primarily due to the limited availability of labeled support examples. This scarcity makes it difficult to capture the full variability of object appearances, leading to poor performance in the presence of occlusions, appearance shifts, and viewpoint differences between support and query samples. To overcome these limitations, we first propose a transductive meta-learning framework that leverages an ensemble of features from pretrained classification and semantic segmentation networks. This method enhances discriminative power by capturing both high-level semantic cues and pixel-level spatial information, and introduces a two-pass correlation mechanism to improve intra-class and intra-object similarity modeling while reducing false positives — all with minimal trainable parameters.

However, despite strong performance, this approach remains limited in its ability to reason about object semantics or adapt flexibly to complex query-support discrepancies. Motivated by these challenges, we introduce a second framework that unifies visual features with semantic knowledge derived from large multimodal language models (LLMs). By generating adaptive class-specific semantic prompts using multi-modal LLMs and integrating them with dense visual correspondences between support and query samples, our model performs reasoning-driven segmentation and achieves robust generalization even in cross-domain setting. The resulting vision-language system addresses key failure cases of prior work, particularly in scenes with severe appearance variation or ambiguous context.

Extensive experiments on Pascal-5^i and COCO-20^i demonstrate that our proposed frameworks outperform prior methods, both in standard few-shot settings and under cross-domain evaluation. Together, these contributions represent a significant advancement in learning to segment with limited supervision, offering a path forward for more intelligent and adaptable vision systems.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Charalambos Poullis, for his unwavering support, insightful guidance, and continuous encouragement throughout my PhD journey. His expertise, vision, and mentorship have been instrumental in shaping my research and academic development. I am especially grateful for the freedom he gave me to explore new ideas and the trust he placed in my work.

I would also like to extend my heartfelt appreciation to my family. To my father and mother, whose endless love, sacrifices, and belief in me have been the foundation of all my achievements. Your encouragement has carried me through the most challenging moments of this journey. To my sister, thank you for your constant support and for always reminding me of the bigger picture.

Finally, I am profoundly grateful to my partner, Houriyeh, for her patience, kindness, and unwavering faith in me. Your presence, love, and emotional support have meant more than words can express, and I could not have completed this journey without you by my side.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Image understanding is a fundamental task in computer vision, typically comprising three primary stages: image classification, object detection, and semantic segmentation. Image classification employs holistic image information to assign an image to one or more predefined categories. Object detection extends this concept by localizing and classifying specific regions of interest (ROIs) within images, typically indicated by bounding boxes. Semantic segmentation represents an even more detailed task, classifying each pixel within an image into predefined classes, thereby providing a fine-grained understanding of image content.

However, deep neural networks' effectiveness typically hinges on access to vast datasets with detailed annotations, which are costly and challenging to acquire, particularly for scenarios involving numerous classes, rare objects, intra-class variability, or tasks requiring specialized and expensive data collection, such as medical or pharmaceutical research Dosovitskiy et al. (2020); K. He, Zhang, Ren, and Sun (2016); Redmon, Divvala, Girshick, and Farhadi (2016); Redmon and Farhadi (2016); Ren, He, Girshick, and Sun (2015). Consequently, approaches like semi-supervised, weakly supervised, zero-shot, and few-shot learning have emerged to address these annotation limitations.

Few-shot learning Finn, Abbeel, and Levine (2017); Snell, Swersky, and Zemel (2017); Vinyals, Blundell, Lillicrap, Wierstra, et al. (2016), in particular, aims to replicate the human ability to quickly learn and recognize new categories from minimal examples. This approach is especially crucial in semantic segmentation, where obtaining pixel-level annotations for training can be exceptionally demanding. Few-shot semantic segmentation (FSS) addresses the task of accurately

segmenting novel classes from a limited number of densely annotated examples Boudiaf, Kervadec, Imtiaz Masud, and Piantanida (2021); Fan, Pei, Tai, and Tang (2022); Hong, Cho, Nam, Lin, and Kim (2021); Lang et al. (2022); W. Liu, Zhang, Lin, and Liu (2020); Okazawa (2022); Shaban, Bansal, Liu, Essa, and Boots (2017); Y. Sun et al. (2022); K. Wang, Liew, Zou, Zhou, and Feng (2019); Z. et al. (2020); C. Zhang, Lin, Liu, Yao, and Shen (2019). This task presents several challenges that critically affect model generalization. A major limitation is feature overfitting—models often extract representations overly tailored to the small support set, leading to poor performance when faced with diverse query images. This issue becomes especially severe in the presence of appearance shifts, such as changes in pose, scale, lighting, or occlusion. For instance, in Pascal-5i, a 'bicycle' may appear clearly in a side profile in the support image but be partially occluded or viewed from a top angle in the query image, significantly impairing segmentation accuracy. Many objects undergo non-rigid transformations or appear partially obscured in real-world scenes, making it difficult for the model to match them accurately. This is particularly common in COCO-20i, where objects like people or animals can appear in varying poses, or hidden behind other objects.

Additionally, false positives—where the model incorrectly classifies regions belonging to base classes or visually similar distractors—are a recurring problem in few-shot setups. Our first paper directly addresses this by integrating a frozen semantic segmentation backbone to suppress base class activations and mitigate misclassification.

To address these challenges, our first paper introduced a transductive meta-learning method motivated by the observation that pretrained classification networks alone, despite their rich semantic clues, provide suboptimal feature representations for segmentation tasks. Additionally, we observed significant performance degradation due to false positives arising from base class features and poor similarity between support and query images. Our approach combined pretrained classification and semantic segmentation networks to leverage their complementary strengths, enhancing feature diversity and discriminative capability. To further address issues such as poor visual similarity between support and query images and the prevalence of false positives, we introduced a two-step transductive segmentation strategy. In the first pass, the model identifies intra-class similarities by matching support foreground features with query features, leveraging a novel ensemble of visual

features from pretrained classification and segmentation networks. This helps overcome discrepancies in appearance due to pose, scale, or occlusion. In the second pass, the model refines its predictions by using the initial query segmentation as support to learn intra-object similarity within the query image itself. This approach enhances feature propagation and suppresses background and base class activations, significantly reducing false positives. Crucially, this method achieves state-of-the-art performance while maintaining minimal trainable parameters, demonstrating an efficient and robust solution to core challenges in FSS.

Building upon these insights, our second paper addressed two key limitations of conventional few-shot segmentation methods: the restricted visual diversity captured by a few support images and the poor generalization resulting from appearance variation between support and query samples. These limitations often lead to biased feature representations and inaccurate segmentation.

To overcome this, we proposed DSV-LFS, a novel framework that leverages the semantic richness of large language models (LLMs) in conjunction with visual matching. First, we use LLMs to generate detailed class descriptions tailored to the object class using carefully crafted prompts. We then introduce a multimodal class semantic encoder that dynamically adapts these descriptions to the specific visual content of the query image, resulting in a query-specific semantic prompt. This semantic prompt captures high-level contextual cues beyond what is available in the support images alone.

In parallel, we introduced a dense visual matching module that performs fine-grained pixel-wise correlation between support and query features to produce a complementary visual prompt. This module captures low-level visual details and spatial correspondences that are critical for precise segmentation. These two prompts—semantic and visual—are then fused within a prompt-based decoder, enabling the network to align contextual class semantics with localized visual evidence in the query image.

This dual-guided segmentation approach significantly improves the model's robustness to occlusions, deformations, and background clutter, as shown through state-of-the-art performance on the Pascal-5i and COCO-20i benchmarks. Unlike earlier LLM-based segmentation methods, which require multi-stage pipelines and post-processing of text outputs, DSV-LFS enables direct, end-to-end segmentation with a single forward pass.

Through extensive experimental evaluations on benchmark datasets such as Pascal-5i and COCO-20i, our dual-prompt framework demonstrated consistent state-of-the-art performance and strong generalization capabilities.

Ultimately, this thesis advances few-shot semantic segmentation by proposing innovative methods that effectively integrate advanced multimodal language models with detailed visual-semantic alignment techniques. These contributions substantially enhance segmentation accuracy, generalization, and robustness, providing a solid foundation for future research in image understanding under limited annotation conditions.

1.1 Contributions

Our technical contributions are:

- In my first paper, we propose a novel transductive, end-to-end framework for few-shot semantic segmentation that addresses three major limitations of existing methods: suboptimal backbone features, poor query-support alignment, and high false positive rates. First, we introduce a multi-backbone feature ensemble that leverages pretrained classification and semantic segmentation networks to combine high-level semantics with pixel-level details. The segmentation backbone also acts as a base-class extractor to suppress false positives from the background. Second, we design a two-pass dense correlation mechanism that learns both intra-class similarity between support and query images, and intra-object similarity within the query itself, enabling effective refinement without additional parameters. Extensive experiments on Pascal-5i and COCO-20i demonstrate that our approach achieves state-of-the-art performance with only 2.98M trainable parameters, outperforming prior methods by significant margins across both 1-shot and 5-shot settings. This work was published as a journal paper in Scientific Reports, 2024 titled "Transductive Meta-Learning with Enhanced Feature Ensemble for Few-shot Semantic Segmentation".

- In my second paper, we present the first approach that integrates large language models (LLMs), fine-tuned for reasoning-based segmentation Lai et al. (2024), with foundation semantic segmentation networks for direct application to few-shot semantic segmentation. We

propose a novel single-stage, end-to-end architecture that effectively combines multimodal semantic cues from LLMs with pixel-level visual features, resulting in improved segmentation accuracy and robustness. Comprehensive experiments on multiple benchmark datasets demonstrate that our method achieves state-of-the-art performance, significantly outperforming existing approaches. This work was published as a conference paper in IEEE/CVF Computer Vision and Pattern Recognition, 2025 titled "DSV-LFS: Unifying LLM-Driven Semantic Cues with Visual Features for Robust Few-Shot Segmentation".

For every study presented in this thesis, I conceived the algorithms, implemented the prototypes, and served as the primary author of the resulting publications, under the guidance of my supervisor.

1.2 Acknowledgements

The work in Chapter 3 was financially supported by the Natural Sciences and Engineering Research Council of Canada Grants RGPIN-2021-03479 (NSERC DG) and ALLRP 571887 - 2021 (NSERC Alliance).

The research in Chapter 4 was undertaken, in part, based on support from the Natural Sciences and Engineering Research Council of Canada Grants RGPIN-2021-03479 (NSERC DG) and ALLRP 571887 - 2021 (NSERC Alliance).

Chapter 2

Preliminary

2.1 Problem Definition

Few-shot semantic segmentation (FSS) addresses the challenge of segmenting a target class in a query image using only a few annotated support examples. To closely mirror real-world deployment conditions, FSS models are trained using an episodic learning framework. Each episode comprises a support set and a query set, simulating the conditions encountered during inference.

In the standard K-shot setting, the support set is defined as

$$S = \{(X_i^s, M_i^s)\}_{i=1}^{K},$$

where X_i^s denotes the i-th support image and M_i^s is its corresponding binary segmentation mask. The query set is represented as

$$Q = (X^q, M^q),$$

where X^q is the query image and M^q is its ground truth segmentation mask used for supervision.

Training episodes are constructed from a predefined set of base (seen) classes, denoted as C_{train}, while evaluation is conducted on a disjoint set of novel (unseen) classes C_{test}, such that

$$C_{\text{train}} \cap C_{\text{test}} = \emptyset.$$

The training dataset for meta-learning is thus defined as

$$D_{\text{train}} = \{(S_i, Q_i, C_i)\}_{i=1}^{N_{\text{train}}},$$

and the test dataset as

$$D_{\text{test}} = \{(S_i, Q_i, C_i)\}_{i=1}^{N_{\text{test}}}.$$

The objective is to train the model on D_{train} in such a way that it can effectively generalize to D_{test}, even though the class labels in the test set were never seen during training.

2.2 Problem Definition with Class Descriptions

In our second paper, we build upon the standard FSS framework by introducing class-level semantic descriptions to enhance the model's understanding of novel classes. These descriptions are automatically generated from the class label C using a large language model (e.g., ChatGPT), enabling the integration of high-level semantic context without introducing external supervision beyond what is available through the support annotations.

Formally, for each target class C, a textual description W_C is generated to characterize its visual and contextual attributes. The training and evaluation datasets are thus augmented as follows:

$$D_{\text{train}} = \{(S_i, Q_i, W_{C_i})\}_{i=1}^{N_{\text{train}}}, \quad D_{\text{test}} = \{(S_i, Q_i, W_{C_i})\}_{i=1}^{N_{\text{test}}}.$$

The semantic description W_C is embedded via a dedicated prompt encoder and injected into the network to complement the visual information extracted from the support set S and the query image X^q. This auxiliary information allows the model to construct a richer and more generalizable representation of the target class, particularly in cases where visual variability or ambiguity hinders performance.

By formulating the task in this manner, we maintain compatibility with conventional FSS protocols while significantly enhancing the robustness and expressiveness of the model through the integration of contextual semantics.

Chapter 3

Transductive Meta-Learning with Enhanced Feature Ensemble for Few-shot Semantic Segmentation

This chapter is a verbatim copy of the journal paper titled "Transductive Meta-Learning with Enhanced Feature Ensemble for Few-shot Semantic Segmentation" authored by A. Karimi, C. Poullis, and published in Scientific Reports, 2024.

Abstract

This paper addresses few-shot semantic segmentation and proposes a novel transductive end-to-end method that overcomes three key problems affecting performance. First, we present a novel ensemble of visual features learned from pretrained classification and semantic segmentation networks with the same architecture. Our approach leverages the varying discriminative power of these networks, resulting in rich and diverse visual features that are more informative than a pretrained classification backbone that is not optimized for dense pixel-wise classification tasks used in most state-of-the-art methods. Secondly, the pretrained semantic segmentation network serves as a base class extractor,which effectively mitigates false positives that occur during inference time and are caused by base objects other than the object of interest. Thirdly, a two-step segmentation approach

using transductive meta-learning is presented to address the episodes with poor similarity between the support and query images. The proposed transductive meta-learning method addresses the prediction by first learning the relationship between labeled and unlabeled data points with matching support foreground to query features (intra-class similarity) and then applying this knowledge to predict on the unlabeled query image (intra-object similarity), which simultaneously learns propagation and false positive suppression. To evaluate our method, we performed experiments on benchmark datasets, and the results demonstrate significant improvement with minimal trainable parameters of $2.98M$. Specifically, using Resnet-101, we achieve state-of-the-art performance for both 1-shot and 5-shot Pascal-5^i, as well as for 1-shot and 5-shot COCO-20^i.

3.1 Introduction

Deep neural networks can learn rich information about visual features of classes that appear in images when trained on vast amounts of labeled data. These attributes significantly contributed to various critical applications, including medical applications Bilal, Sun, Mazhar, Imran, and Latif (2022); Bilal, Zhu, Deng, Lu, and Wu (2022). However, their ability to generalize to new classes diminishes when presented with only a limited number of labeled examples Z. Li, Kamnitsas, and Glocker (2019), which is a prevalent issue in domains such as geospatial and medical, where collecting and labeling large datasets is a complex and expensive process. To overcome this issue, researchers have proposed the few-shot learning paradigm, which attempts to mimic the capacity of the human visual system to rapidly learn new classes from a small number of labeled examples.

This paper focuses on few-shot semantic segmentation, a special case of semantic segmentation in which the model must generalize to novel(unseen) classes and classify the pixels in an image. The most challenging aspect of few-shot segmentation is fully utilizing the information in the small support set of training examples K on N unseen classes (N-way, K-shot for $K < 5$). Two primary strategies for fewshot image understanding are proposed. The first strategy centers on the learning-to-learn (or meta-learning) paradigm. In order to simulate the tasks that will be presented during inference, meta-learning strategies popularized the necessity of organizing training data into episodes B., J., and T. (2021); Fan et al. (2022); H., D., K., S., and Y. (2021); Hong et al. (2021);

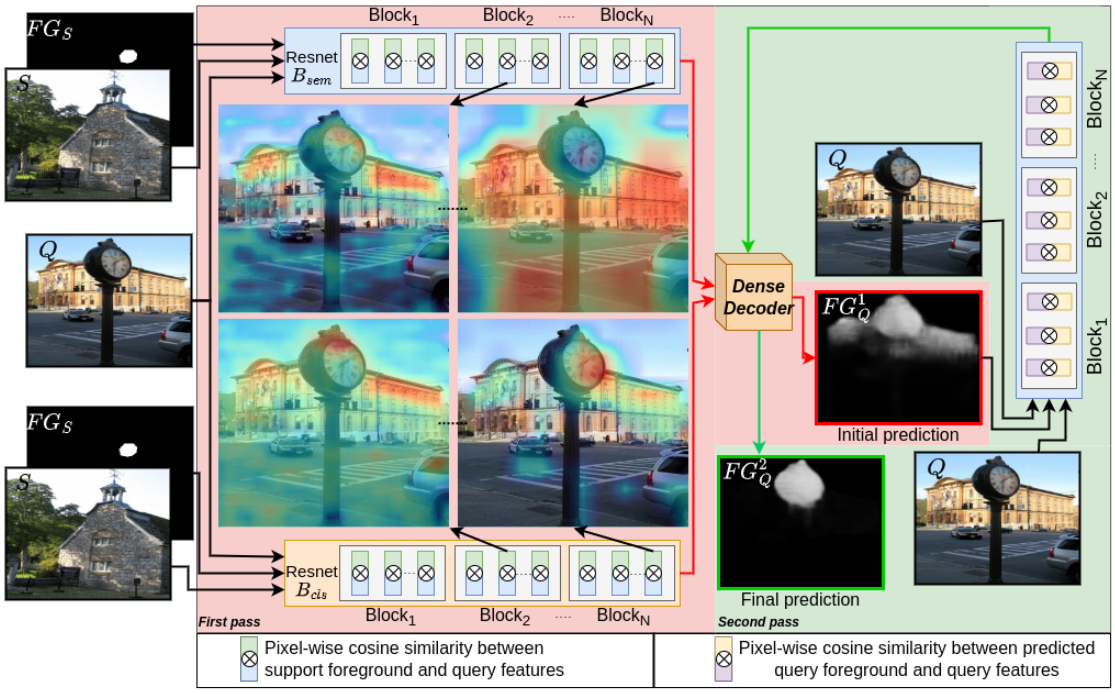


Figure 3.1: We propose two-pass end-to-end method for few-shot semantic segmentation. The approach leverages an ensemble of visual features learned from pretrained classification B_{cls} and semantic segmentation B_{sem} networks with the same architecture. B_{sem} is also used as a base class extractor. The first pass (red background) matches support foreground features to query features to address intra-class variation, and the second pass (green background) suppresses false positives and propagates query foreground features to leverage intra-object variation. Heatmaps show pixel-correlations between the query features and support foreground features in different layers of the network

Lang et al. (2022); Y. Liu, Zhang, Zhang, and He (2020); Okazawa (2022); Shi et al. (2022); Y. Sun et al. (2022); K. Wang et al. (2019); B. Yang, Liu, Li, Jiao, and Ye (2020); Z. et al. (2020); C. Zhang et al. (2019). Similar to standard training, the second line of research addresses few-shot image understanding by training a network using base classes and fine-tuning with novel classes Boudiaf et al. (2021); W.-Y. Chen, Liu, Frank Wang, and Huang (2019); Dhillon, Chaudhari, Ravichandran, and Soatto (2019); Guo et al. (2020); Y. Liu, Lee, and Park (2019); Masud Ziko, Dolz, Granger, and Ben Ayed (2020); Rodrıguez, Laradji, Drouin, and Lacoste (2020); Y. Tian, Wang, Krishnan, and Tenenbaum (2020). A frozen pretrained classification backbone is utilized by the both line of researches because it has been demonstrated to generalize more effectively to unseen classes.

The first observation is that a pre-trained classification backbone on a large-scale dataset such

as Image-Net contains rich semantic clues; however, it is suboptimal to adopt directly for a segmentation task Y. Sun et al. (2022). Nevertheless, the majority of recent techniques have shown that fine-tuning a pre-trained classification backbone during the episodic training phase is susceptible to overfitting. Experiments presented in Y. Sun et al. (2022) to fine-tune the entire backbone or a subset of layers in FSS demonstrate a negative effect on the final result. Consequently, during the episodic phase, updating millions of backbone parameters necessitates careful training considerations and increases the demand for training resources and time. The recent work to address this issue Y. Sun et al. (2022), which achieves state-of-the-art with a Resnet-50, significantly increases the memory requirements compared to other few-shot techniques. The objective of updating the backbone is to provide enhanced pixel-level features, which is an open problem in FSS. To achieve enhanced pixel-level features without fine-tuning the backbone, we investigated the distinctions between a classification and segmentation backbone. Classification networks are trained with image-level labels and learn visual features that incorporate the spatial distribution and shape of the objects at a higher level of abstraction. In contrast, semantic segmentation networks trained on pixel-level labels discover visual features at the pixel-level that incorporate contextual information based on the spatial relationships between different objects in the image Badrinarayanan, Handa, and Cipolla (2015); Long, Shelhamer, and Darrell (2015); Noh, Hong, and Han (2015); Zeiler and Fergus (2014). In other words, the discriminative power of a semantic segmentation network is higher at intermediate layers, while a classification network has a higher discriminative power at the final layers. We present the experiments and analysis on the impact that a pretrained backbone can have on the pixel-wise feature correlations, when it is pretrained on a classification versus a semantic segmentation task. For the comparison, we used a frozen classification backbone pretrained on ImageNet-1K and a frozen semantic segmentation backbone pretrained on base classes as described in Lang et al. (2022). We analyzed the pixel-correlations between the query features and support foreground features by calculating the discriminative power of the features at each backbone layer. The discriminative power ρ^k at layer k is measured as the ratio $\rho^k = \frac{\frac{1}{N}\sum_i^N cos(FG_Q^i, P_S)}{\frac{1}{M}\sum_j^M cos(BG_Q^j, P_S)}$ where P_S is the support prototype calculated by averaging all the foreground support features FG_S. The numerator is the average cosine distance of the N foreground query features $FG_Q^i, 0 \leq i \leq N$ to the foreground support prototype FG_S, and the denominator is the average cosine distance of

the M background query features $BG_Q^j, 0 \leq j \leq M$ to FG_S. Intuitively, the higher the ratio ρ^k the higher the discriminative power to differentiate between the query foreground and background features w.r.t. the support foreground features at layer k. Figures 3.2a and 3.2b show the discriminative power calculated using more than $4,000$ episodes from Pascal-5^i, ρ^k of each backbone at layers $k, 1 \leq k \leq |B_{cls}|$ of the frozen pretrained backbones B_{cls} and B_{sem}. Figure 3.2c top-left, shows the query image, with an inset of the corresponding support image. The remaining panels depict pixel-correlations between the query features and support foreground features in different layers (from left to right, intermediate layers to final layers) of a pretrained classification network (top row) and a semantic segmentation network (bottom row) which shows the higher discriminative power of the semantic segmentation network in the intermediate layers and similarly, for the classification network in the final layers. Utilizing the advantages of each, we present a multi-scale feature ensemble comprised of visual features learned by pretrained classification and semantic segmentation networks to specifically satisfy the need for both rich semantic cues and pixel-level information.

The second observation is that the object in the support image is frequently not visually similar to that in the query image. Several factors contribute to this, including viewpoint variation, illumination changes, scale, deformation, occlusion, intra-class variation, clutter, and motion. Consequently, query segmentation may contain some errors. Numerous techniques for self-refinement based on initial query prediction have been proposed B. et al. (2021); Fan et al. (2022); Min, Kang, and Cho (2021); K. Wang et al. (2019). Recently, Fan et al. (2022) proposed a two-step segmentation method by utilizing the high confidence area of initial query prediction via non-differentiable thresholding, which has a number of limitations. In contrast, we present a two-pass end-to-end dense correlation learning method that enables the network to learn the visual disimilarities between the query foreground features and the false positives without introducing non-differentiable operations or additional components. In the first pass, intra-class similarity is addressed by matching support foreground features to query features, and in the second step, intra-object similarity is addressed by suppressing false positives from the initial prediction and propagating query foreground features throughout the query image. The proposed method does not introduce any additional trainable parameters to the network, whereas the Fan et al. (2022) fine-tunes the last two blocks of a backbone. Moreover, our self-refinement module can operate on top of any backbone, which is another

significant advantage over Fan et al. (2022) which reshapes embedding space for self-refinement.

The third observation is that false positives account for a substantial proportion of incorrect classifications and significantly hinder performance. As noted by Lang et al. (2022), the presence of base classes in the background of the query image can lead to false positive predictions, as the network may incorrectly classify pixels that are not part of the object of interest. To address this issue, they proposed auxiliary layers on top of a base learner that is trained on base classes to predict whether or not each pixel in the output of the meta learner corresponds to a base class. By using this information to selectively mask out base class predictions, they were able to reduce the number of false positives and improve segmentation accuracy. Inspired by this work and based on observations from our extensive experimentation -as described in the appendix- we propose a method that reduces false positives caused by base classes that is both simpler and faster than the method proposed by Lang et al. (2022), resulting in a shorter training time with the same functionality and performance.

In this paper, we present a two-pass end to end method for few-shot semantic segmentation that addresses each of the aforementioned issues. The proposed method (Figure 3.1) leverages an ensemble of visual features learned by segmentation and classification backbones to segment a query image in two steps. Dense convolutional layers trained to match support objects to query images using ensemble features in first step and propagate initial query predictions in the second step.

We evaluate our method on the benchmark datasets Pascal-5^i and COCO-20^i, and report our results. On Pascal-5^i 1-shot and 5-shot, with a Resnet-101 backbone, we achieve state-of-the-art by a margin of 2.51% and 1.12%, respectively. Similarly, on COCO-20^i 1-shot and 5-shot, with a Resnet-101 backbone, we achieve state-of-the-art by a margin of 3.98% and 1.6%, respectively. Our model has a minimal number of trainable parameters i.e. $2,980,711$ compared to the baseline Min et al. (2021) i.e. $2,587,394$.

The paper is organized as follows: Sec. 3.2 outlines the most recent and pertinent work in few-shot semantic segmentation. In Sec. 3.3, the methodology is described in depth, followed by the experiments and ablations on the two benchmark datasets Pascal-5^i and COCO-20^i in Sec. 3.4. We conclude and provide suggestions for future work in Sec. 3.5.

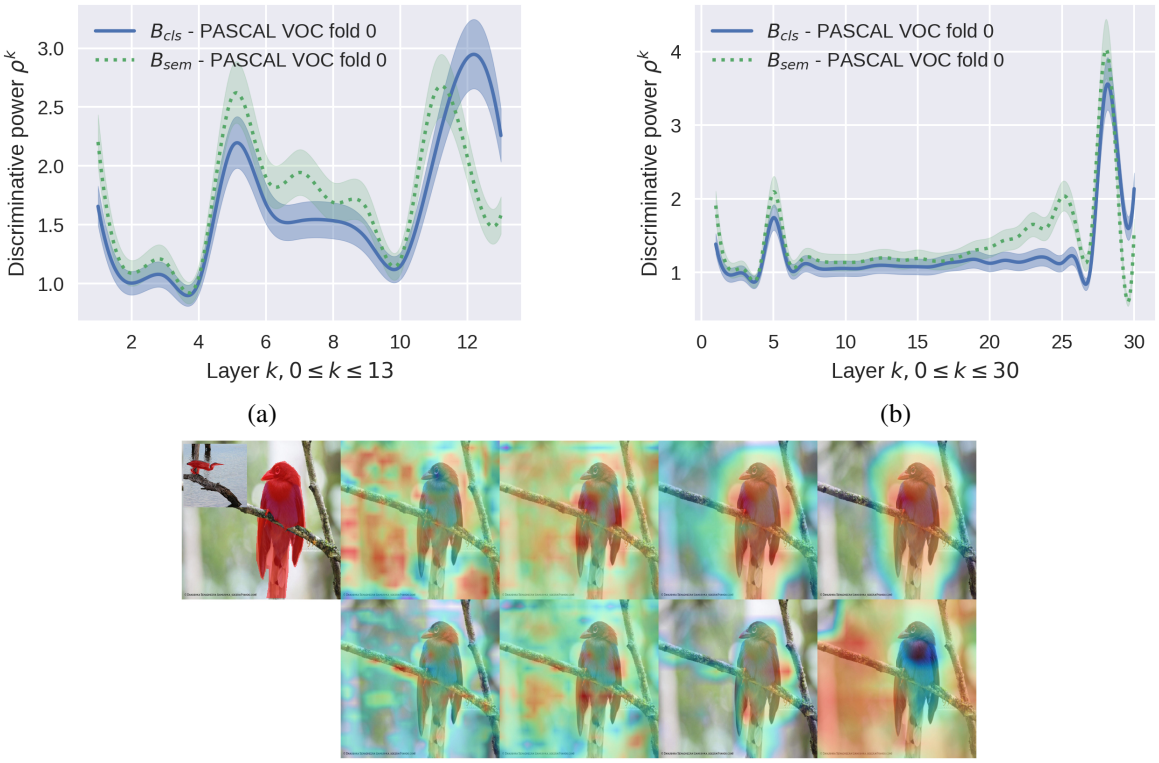
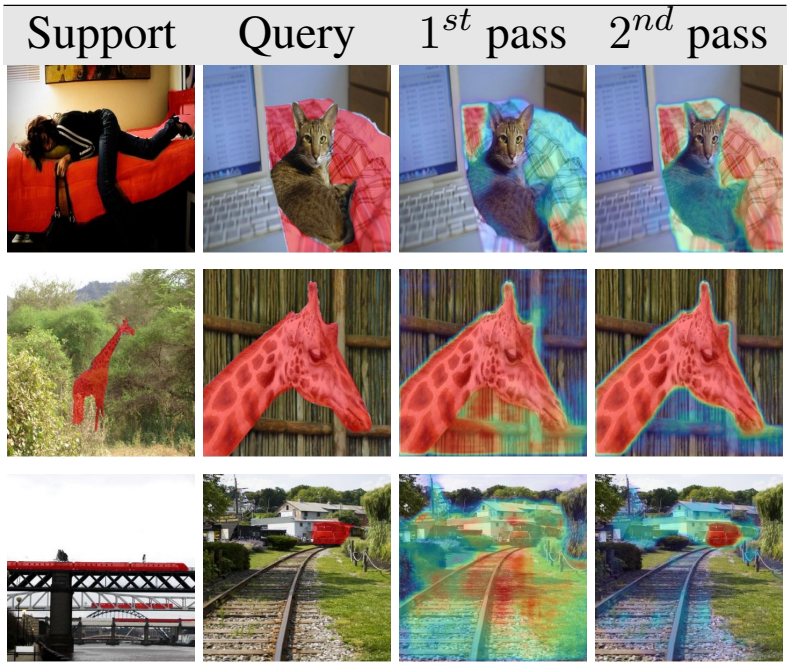


Figure 3.2: **Discriminative power of classification vs semantic segmentation networks.** (a) ——: classification network (Resnet-50), - - -: semantic segmentation network (Resnet-50). The discriminative power ρ^k at layer k is measured as the ratio $\rho^k = \frac{\frac{1}{N}\sum_i^N cos(FG_Q^i, P_S)}{\frac{1}{M}\sum_j^M cos(BG_Q^j, P_S)}$ (b) Same as (a) but for Resnet-101. Graphs for all folds are in the appendix. (c) The top left shows the query image, with an inset of the corresponding support image. The remaining panels depict pixel-correlations between the query features and support foreground features in different layers (from left to right, intermediate layers to final layers) of a pretrained classification (top row) and semantic segmentation networks (bottom row). The discriminative power of a semantic segmentation network is higher at intermediate layers, and the discriminative power of a classification network is higher at the final layers as also demonstrated in (a).

3.2 Related Work

Few-shot learning techniques enable learners to generalize to new classes using a small number of labeled samples. These techniques follow a similar pipeline: a pre-trained backbone network is used to generate feature embeddings from input images, and a model head is used to generate segmentation maps using these embeddings as input. Numerous techniques have been proposed that fall into one of four broad categories: (i) metric learning techniques where the objective is to learn

Table 3.1: **Results from our two-pass method.** 1^{st} pass: intra-class similarity ($S \longrightarrow Q$). 2^{nd} pass: intra-object similarity ($Q \longrightarrow Q$).



a mapping from image space to feature space that ensures the distance between feature vectors of similar categories is small, while it is large for dissimilar categories Gidaris and Komodakis (2018); T. Hu et al. (2019); Satorras and Estrach (2018); Vinyals et al. (2016); K. Wang et al. (2019), (ii) initialization-based techniques where the objective is to learn a good model initialization so that fine-tuning is possible with a few training examples and a small number of gradient update steps Finn et al. (2017); W. Liu et al. (2020); Rakelly, Shelhamer, Darrell, Efros, and Levine (2018a, 2018b); Ravi and Larochelle (2016); Rusu et al. (2018); Snell et al. (2017); Sung et al. (2018), (iii) Hallucination-based techniques where the objective is to learn a generator from the available data that "hallucinates" novel class data for data augmentation Hariharan and Girshick (2017); Y.-X. Wang, Girshick, Hebert, and Hariharan (2018), (iv) semantic-based learning techniques where the objective is to learn a generator conditioned on additional attributes, typically semantic word embeddings. Then, a layer for fine-tuning classification is applied to features from both types of classes. Bucher, Tuan-Hung, Cord, and Pérez (2019); A. Li, Luo, Lu, Xiang, and Wang (2019); Schwartz, Karlinsky, Feris, Giryes, and Bronstein (2019).

Our work falls into the metric-based techniques and is trained with episodic training as proposed

by initialization-based approaches. Early work with metric-based approaches used a two branch network to find the most similar area in the query image using extracted support prototypes based on distance measures, such as Euclidean distance and cosine distance K. Wang et al. (2019). Other work proposed additional modules to compare query pixels and support prototypes C. Zhang et al. (2019), while others focused on the limited representation capability of a single prototype and proposed methods to develop multiple prototypes to perform comparisons B. et al. (2021); H. et al. (2021); Y. Liu et al. (2020); B. Yang et al. (2020).

Recently, W.-Y. Chen et al. (2019); Dhillon et al. (2019); Guo et al. (2020); Masud Ziko et al. (2020); Y. Tian et al. (2020) reevaluated the use of cross-entropy for training the network on base classes and demonstrated that competitive performance could be attained through fine-tuning on unseen classes. Following this pattern, works including Dhillon et al. (2019); Y. Liu et al. (2019); Rodrıguez et al. (2020) demonstrated that transductive few-shot learning could enhance performance. Specifically, Boudiaf et al. (2021) attained competitive performance by incorporating transductive loss terms into the training and then fine-tuning a single classifier layer trained on base classes. Shannon entropy Boudiaf et al. (2021); Dhillon et al. (2019) on each query sample and KL divergence on background/foreground distribution of samples Boudiaf et al. (2021) are the two most common transductive losses. This research demonstrated that transductive learning could not generalize to a new class, however, it can learn the characteristics of a specific sample of a new category, substantially improving the final results.

Currently, few-shot semantic segmentation techniques tend to use all available information and learn the visual similarities between the pixels in the query and support image. Particularly, all-pairs field transforms introduced by Teed and Deng (2020) for visual similarities contributed to further considerable gains in few-shot semantic segmentation. The authors of Min et al. (2021) recast few-shot semantic segmentation as a visual similarity task and perform N^4 all-pairs visual comparisons between the pixel-level features in the query and support images. Instead of learning similarities between class prototypes, their network is trained on the visual similarities between all pixel pairings at various network layers.

Several methods have recently shown a considerable performance improvement using pre-trained transformer backbone. Shi et al. (2022) suggested a method for computing similarities between

query pixels and all support pixels using a multi-level pixel-wise attention module. The authors reported a substantial improvement when employing a pre-trained transformer backbone as opposed to a convolutional backbone such as ResNet. J. Zhang, Sun, Yang, and Chen (2022) revives the framework of using a backbone for feature extraction followed by a linear classification head. The authors propose a transformer as the backbone and a classification head that combines pixel-level and class-level features, which has been shown to capture global context better than a convolutional network, significantly boosting performance. Recent works such as Shi et al. (2022), have demonstrated significant gains in performance, however, this can easily be attributed to the vision transformer backbone rather than the effectiveness of their proposed technique.

Despite these advancements, there are still challenges to overcome, most notably the bias towards the base classes and insufficient visual similarity between the support and query image, which can result in subpar performance. Our method alleviates these issues. Specifically, our technical contributions are as follows:

- An end-to-end transductive learning method for few-shot semantic segmentation. Uniquely, the matching operates on dense, multi-level visual similarities between support-query pixels and query-query pixels in the first and second passes, respectively, as shown in Table 3.1.
- A feature ensemble comprised of visual features learned by pretrained classification and semantic segmentation networks. Furthermore, the semantic segmentation network, through the use of a simple and efficient pipeline, serves as a base class and background extractor, drastically reducing false positives.
- Our method, using Resnet-101 backbone, achieves state-of-the art performance on 1-shot and 5-shot Pascal-5^i as well as COCO-20^i while requiring only $2.98M$ in trainable parameters.

3.3 Methodology

The input is a pair of images of the same class S and Q which form the support and query respectively.

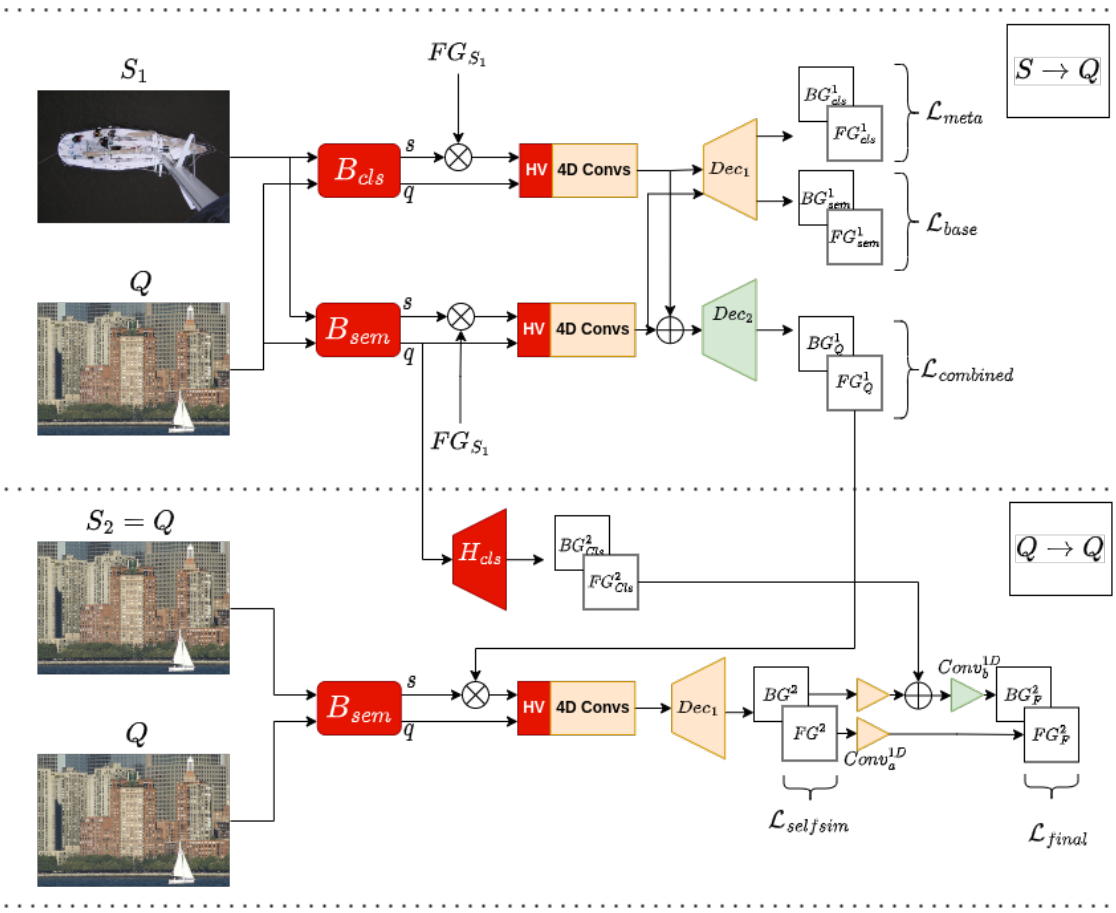


Figure 3.3: **Technical overview of proposed meta-learner.** B_{cls}, B_{sem}: pretrained classification and semantic segmentation networks, respectively (frozen), H_{Cls}: pretrained classification layer (frozen), **HV**: Hypercorrelation volumes (multi-scale cosine similarity between features with no trainable parameters), **4D Convs**: 4D convolutions resulting in correlation tensors in $\mathbb{R}^{C\times H\times W\times H\times W}$ for feature tensors with dimensions $C\times H\times W$, followed by concatenation across scale and an average pooling on the last two dimensions to reduce the dimensions to $\mathbb{R}^{C\times H\times W}$, ▷, ▶: 1D Convolution; the first two $Conv_a^{1D}$ share weights, **BG/FG**: Background/Foreground, **Dec**: a decoder; Decoders shown in yellow are the same. Red indicates a frozen module, Orange indicates shared trainable parameters, and Green indicates a module with individually trainable parameters. Total number of trainable parameters: $2.98M$.

3.3.1 Learning intra-class similarity $S \rightarrow Q$

The first pass takes the support $S_1 = S$ and query Q as inputs. The objective of this pass is to learn intra-class similarity by learning features from the support S_1 and segment visually similar features in the query Q.

A backbone B_{cls} is a frozen pretrained classification network that learns features $f^{cls}_{S_1}$ and f^{cls}_{Q}

for image S_1 and Q, respectively. These features encode the spatial distribution and shape of the objects at a more abstract level. This information is supplemented by features $f_{S_1}^{sem}$ and f_Q^{sem} learned by a backbone B_{sem}, a frozen semantic segmentation network trained on both background and base classes. The training of B_{sem} with pixel-level labels results in features $f_{S_1}^{sem}$ and f_Q^{sem} capturing contextual information and spatial similarities.

Support features from the two backbones, $f_{S_1}^{cls}$ and $f_{S_1}^{sem}$, are multiplied by the foreground mask FG_{S_1} in order to remove background-related features.

$$f_{S_1}^{cls} = FG_{S_1} \otimes B_{cls}(S_1), f_Q^{cls} = B_{cls}(Q) \tag{1}$$

$$f_{S_1}^{sem} = FG_{S_1} \otimes B_{sem}(S_1), f_Q^{sem} = B_{sem}(Q) \tag{2}$$

Next, we compare the support and query features by computing the cosine similarity between all pairs of pixels in $f_{S_1}^i$ and f_Q^i, where $i \in \{cls, sem\}$. This is performed at different depths of each backbone leading to a set of multi-scale 4D volumes, each given by,

$$HV(f_S^i, f_Q^i) = \text{ReLU}\left(\frac{f_S^i \cdot f_Q^i}{\left|f_S^i\right| \cdot \left|f_Q^i\right|}\right), \tag{3}$$

where $i \in \{cls, sem\}$. For features with dimensions in $\mathbb{R}^{C \times H \times W}$ the dimensions of the volume are $\mathbb{R}^{C \times H \times W \times H \times W}$, where C is the number of channels, and H, W are the height and width, respectively. This module does not have any trainable parameters.

4D convolutions are applied on the set of multi-scale hypercorrelation volumes. This module, adapted from Min et al. (2021), applies the 4D convolutions on center-pivot pixels to reduce the memory and time requirements. Incrementally, lower scale features are upsampled and concatenated with higher scale features, followed by average pooling on the last two dimensions in order to reduce the dimensions of the concatenated correlations $\mathcal{C}_{cls}$ and $\mathcal{C}_{sem}$ to $\mathbb{R}^{C \times H \times W}$.

$$\mathcal{C}_{cls} = AvgPool(Conv^{4D}(HV(f_{S_1}^{cls}, f_Q^{cls}))) \tag{4}$$

$$\mathcal{C}_{sem} = AvgPool(Conv^{4D}(HV(f_{S_1}^{sem}, f_Q^{sem}))) \tag{5}$$

The first pass concludes with two decoders Dec_1 and Dec_2. Dec_1 operates on $\mathcal{C}_{cls}$ and $\mathcal{C}_{sem}$ and for each generates a semantic segmentation mask of the foreground FG_i and background BG_i, where $i \in \{cls, sem\}$, supervised with the losses $\mathcal{L}_{cls} = \frac{1}{N}\sum_{n=1}^{N} CE(BG_{cls} \oplus FG_{cls}, Q_n^{gt})$ and $\mathcal{L}_{sem} = \frac{1}{N}\sum_{n=1}^{N} CE(BG_{sem} \oplus FG_{sem}, Q_n^{gt})$, respectively, where Q_n^{gt} is the n^{th} ground-truth query foreground mask, $n \in N$. Dec_2 transforms the concatenated correlations into FG_Q^1 and BG_Q^1, supervised by loss $\mathcal{L}_{combined}$.

$$FG_{cls}, BG_{cls} = Dec_1(\mathcal{C}_{cls}) \tag{6}$$

$$FG_{sem}, BG_{sem} = Dec_1(\mathcal{C}_{sem}) \tag{7}$$

$$FG_Q^1, BG_Q^1 = Dec_2(\mathcal{C}_{cls} \oplus \mathcal{C}_{sem}) \tag{8}$$

where the superscript $(.)^1$ indicates an outcome of the first pass. The loss is given by $\mathcal{L}_{combined} = \frac{1}{N}\sum_{n=1}^{N}[CE(BG_Q^1 \oplus FG_Q^1, Q_n^{gt}) - \kappa\mathcal{L}_{Sh}]$, where $\kappa = 0.1$. The second term of $L_{combined}$ is the transductive loss term given by Shannon entropy $\mathcal{L}_{Sh}$,

$$\mathcal{L}_{Sh} = \frac{1}{H \times W}\sum_{p=1}^{H \times W}(BG_Q^1(p) \oplus FG_Q^1(p))log(BG_Q^1(p) \oplus FG_Q^1(p)) \tag{9}$$

where $p \in H \times W$ is pixel. The Shannon entropy encourages the network to have a polarised initial prediction with a high or low confidence area S. Dhillon, Chaudhari, and Ravichandran (2020), which reduces the number of false positives. The impact of transductive terms is explained further in the appendix.

3.3.2 Learning intra-object similarity $Q \rightarrow Q$

As input for the second pass, the query image Q serves as both the support $S_2 = Q$ and query Q. The objective of the second pass is to learn intra-object similarity by propagating in the query image Q those features in Q that were visually similar to the features of the support S_1 in the first pass. As mentioned previously, the premise is twofold: (i) that intra-object similarity, which is the visual similarity between features in the same image, is greater than intra-class similarity, which is the visual similarity between features in two different images of the same class, and (ii) that learning

features of the background and base classes reduces false positives. It has been demonstrated that the affinity between unlabeled samples has a significant effect on transductive learning Y. Liu et al. (2019). We observed that a pretrained semantic segmentation backbone has greater pixel affinity than a pretrained classification network. In the second pass, we therefore employ a semantic segmentation backbone.

Features $f^2_{S_2}$ and f^2_Q are extracted by the semantic segmentation backbone B_{sem}. Support features $f^2_{S_2}$ are multiplied by the foreground mask of Q resulting from the first pass. Similar to the first pass, multi-scale hypercorrelation volumes are calculated followed by multi-scale 4D convolutions and average pooling on the last two dimensions. A decoder Dec_1 maps the correlations $\mathcal{C}^2$ into FG^2 and BG^2 segmentation maps which are supervised with the loss $\mathcal{L}_{selfsim} = \frac{1}{N}\sum_{n=1}^{N} CE(FG^2 \oplus BG^2, Q^{gt}_n)$. Each segmentation is then passed through 1D-convolutions sharing weights (shown as ▷ in Figure 3.3).

$$f^2_{S_2} = FG^1_Q \otimes B_{sem}(S_2), f^2_Q = B_{sem}(Q) \tag{10}$$

$$\mathcal{C}^2 = AvgPool(Conv^{4D}(HV(f^2_{S_2}, f^2_Q))) \tag{11}$$

$$FG^2, BG^2 = Dec_1(\mathcal{C}^2) \tag{12}$$

where the superscript $(.)^2$ indicates an outcome of the second pass.

The semantic segmentation backbone B_{sem}, which has been pretrained on background and base classes, serves to eliminate false positives from the query foreground segmentation mask. A pretrained classification layer H_{cls} acting on the backbone's B_{sem} query features $(f^{sem}_Q)^1$ from the first pass, generates foreground FG^1_{Cls} and background BG^1_{Cls} maps. The foreground map FG^1_{Cls} of the classifier contains base classes. In the penultimate step, the background map of B_{sem}, BG^2, is combined with the foreground map of the classifier FB^1_{Cls}, passed through a 1D convolution and finally combined with the foreground probabilities of B_{sem}, FG^2. The final map is supervised with

loss $\mathcal{L}_{final} = \frac{1}{N}\sum_{n=1}^{N} CE(FG_F^2 \oplus BG_F^2, Q_n^{gt})$.

$$FG_{Cls}^1, BG_{Cls}^1 = H_{cls}((f_Q^{sem})^1) \tag{13}$$

$$FG_F^2 = Conv_a^{1D}(Conv_b^{1D}(BG^2) \oplus FG_{Cls}^1) \tag{14}$$

$$BG_F^2 = Conv_b^{1D}(FG^2) \tag{15}$$

The proposed meta-learner (Figure 3.3) is trained using episodic training supervised by the loss L given by,

$$L = L_{cls} + L_{sem} + L_{combined} + L_{selfsim} + L_{final} \tag{16}$$

with equal weights for each term.

Extension to K-shot setting. For K-shot setting, we employ the method in Min et al. (2021). Given K support image-mask pairs and a query image, we perform K forward passes to predict K masks. Voting is conducted at each pixel location by summing the K predictions and dividing each output score by the maximum votes. A pixel is designated as foreground if its voting score exceeds a predetermined threshold.

3.4 Experiments

3.4.1 Implementation details

Modules. The backbones B_{cls} and B_{sem} are frozen Resnet-style backbones pretrained using supervized classification leanirng on ImageNet-1K and supervized segmentation learning on base classes of each fold respectively. The 4D convolutions all share the same architecture and weights, and have $2.5M$ trainable parameters. There are two decoders having the same architecture. We use episodic training to train the meta-learner with the two frozen backbones B_{cls} and B_{sem}.

Training. The training consists of two phases: pretraining and meta-training. Following Lang et al. (2022), we trained a supervised segmentation model on base classes associated with each fold in the first phase. PSPNet with two different backbones, namely ResNet50 and ResNet101, is used as a segmentation model, and it is trained on Pascal-5^i for 100 epochs and COCO-20^i for 20

epochs, with batch size set to 12 and a stochastic gradient descent optimizer with an initial learning rate $2.5e-3$. In the second phase, the entire model with frozen backbones is trained with episodic learning. In the majority of previous FSS methods, it has been demonstrated that frozen backbone facilitate generalisation in episodic learning. For the Pascal-5^i and COCO-20^i, the batch size is set to 24 and 48, respectively, and the model is trained for 200 iterations using the Adam optimizer and an initial learning rate of $1e-3$. No data augmentation is used during training to ensure a fair comparison with other methods. Four NVIDIA V100 GPUs are employed for training.

Table 3.2: **Comparison with current state-of-the-art for Pascal-5^i 1-shot and 5-shot tasks.** The highest values are in **bold**, and the second-highest are underlined. Average mIoU is Highlighted. See appendix for full-sized table.

		1-shot					5-shot				
Backbone	Method	f0	f1	f2	f3	mIoU	f0	f1	f2	f3	mIoU
Resnet-50	REPRIBoudiaf et al. (2021)	60.2	67.0	61.7	47.5	59.1	64.5	70.8	71.7	60.3	66.8
	PFENetZ. et al. (2020)	61.7	69.5	55.4	56.3	60.8	63.1	70.7	55.8	57.9	61.9
	ProtRelOkazawa (2022)	65.2	72.9	63.3	61.3	65.7	70.2	75.6	68.9	66.2	70.2
	VATHong et al. (2021)	67.6	72.0	62.3	60.1	65.5	72.4	73.6	68.6	65.7	70.1
	SSPFan et al. (2022)	60.5	67.8	66.4	51.0	61.4	67.5	72.3	75.2	62.1	69.3
	DCAMAShi et al. (2022)	67.5	72.3	59.6	59.0	64.6	70.5	73.9	63.7	65.8	68.5
	BAM + SVFY. Sun et al. (2022)	**69.38**	**74.51**	**68.80**	63.09	**68.95**	**72.05**	**76.17**	**71.97**	**68.91**	**72.28**
	BAMLang et al. (2022)	68.97	73.59	67.55	61.13	67.81	70.59	75.05	70.79	67.20	70.91
	Baseline-HSNet	64.3	70.7	60.3	60.5	64.0	70.3	73.2	67.4	67.1	69.5
	Ours	68.03	73.69	64.25	**64.72**	67.67	71.26	75.13	67.75	68.11	70.56
Resnet-101	REPRIBoudiaf et al. (2021)	59.6	68.6	62.2	47.2	59.4	66.2	71.4	67.0	57.7	65.6
	PPNetY. Liu et al. (2020)	52.7	62.8	57.4	47.7	55.2	60.3	70.0	69.4	60.7	65.1
	PFENetZ. et al. (2020)	60.5	69.4	54.4	55.9	60.1	62.8	70.4	54.9	57.6	61.4
	ProtRelOkazawa (2022)	67.8	74.6	65.7	62.2	67.5	70.0	75.9	**71.8**	65.8	70.9
	VATHong et al. (2021)	70.0	72.5	64.8	64.2	67.9	**75.0**	75.2	68.4	69.5	72.0
	SSPFan et al. (2022)	60.5	67.8	66.4	51.0	61.4	67.5	72.3	75.2	62.1	69.3
	DCAMAShi et al. (2022)	65.4	71.4	63.2	58.3	64.6	70.7	73.7	66.8	61.9	68.3
	Baseline-HSNet	67.3	72.3	62.0	63.1	66.2	71.8	74.4	67.0	68.3	70.4
	Ours	**71.25**	**76.19**	**67.73**	**66.47**	**70.41**	73.85	**77.53**	70.72	**70.41**	**73.12**

3.4.2 Evaluation

Benchmark datasets. We evaluate the performance of the proposed method on two major few-shot segmentation datasets, Pascal-5^i and COCO-20^i, which were constructed from PASCAL VOC 2012 with 20 classes and MS-COCO datasets with 80 classes, respectively. COCO-20^i is more challenging because it has more classes, samples, and more object instances per image. With minor

Table 3.3: **Comparison with current state-of-the-art for COCO-20^i 1-shot and 5-shot tasks.** The highest values are in **bold**, and the second-highest are underlined. Average mIoU is Highlighted. See appendix for full-sized table.

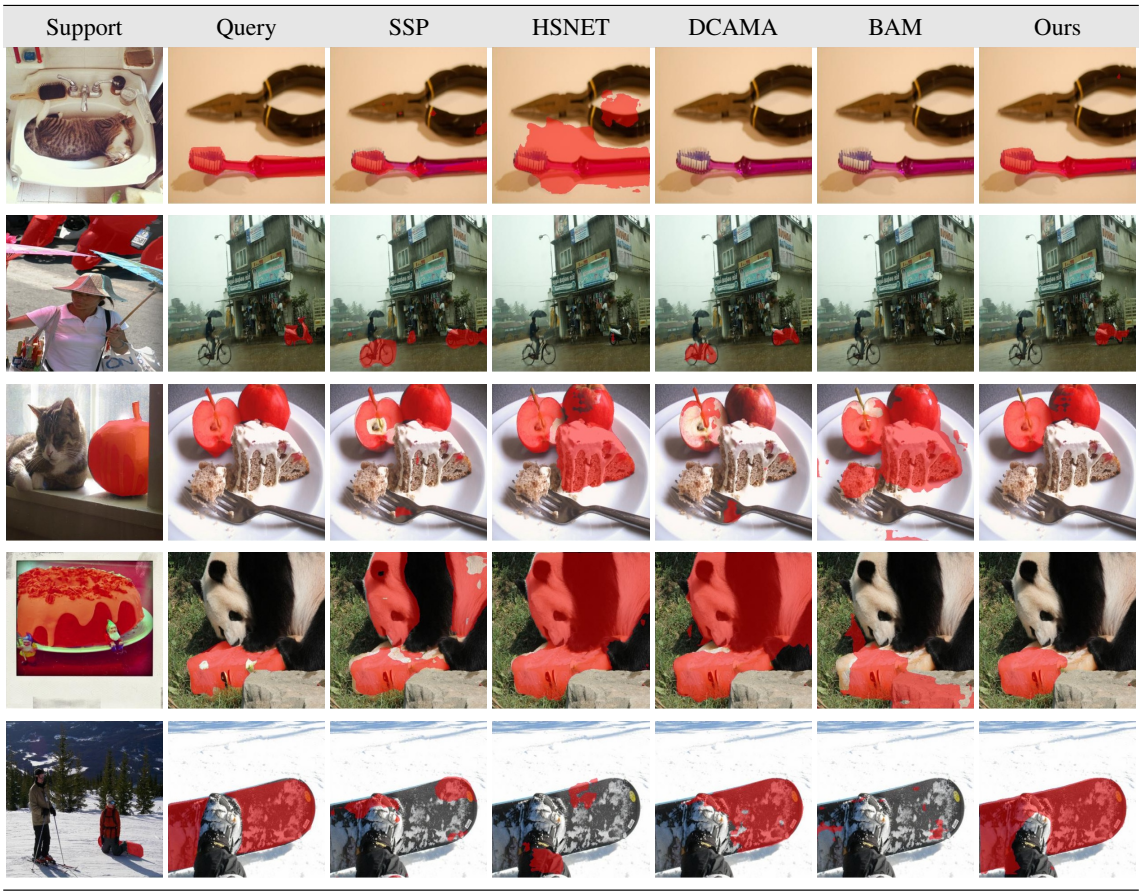
		1-shot					5-shot				
Backbone	Method	f0	f1	f2	f3	mIoU	f0	f1	f2	f3	mIoU
Resnet-50	REPRIBoudiaf et al. (2021)	32.0	38.7	32.7	33.1	34.1	39.3	45.4	39.7	41.8	41.6
	PFENetZ. et al. (2020)	36.5	38.6	34.5	33.8	35.8	36.5	43.3	37.8	38.4	39.0
	ProtRelOkazawa (2022)	42.2	48.9	45.5	44.6	45.3	48.0	55.7	50.7	50.1	51.1
	VATHong et al. (2021)	39.0	43.8	42.6	39.7	41.3	44.1	51.1	50.2	46.1	47.9
	SSP Fan et al. (2022)	35.5	39.6	37.9	36.7	37.4	40.6	47.0	45.1	43.9	44.1
	DCAMAShi et al. (2022)	41.9	45.1	44.4	41.7	43.3	45.9	50.5	50.7	46.0	48.3
	BAM + SVFY. Sun et al. (2022)	**46.87**	**53.80**	48.43	44.78	**48.47**	**52.25**	57.83	51.97	**53.41**	**53.87**
	BAMLang et al. (2022)	43.41	50.59	47.49	43.42	46.23	49.26	54.20	51.63	49.55	51.16
	Baseline-HSNet	36.3	43.1	38.7	39.2	39.2	43.3	51.3	48.2	45.0	46.9
	Ours	42.15	53.22	**49.05**	**48.08**	**48.12**	47.50	**59.14**	**53.19**	51.16	52.75
Resnet-101	PPNetY. Liu et al. (2020)	17.0	18.0	21.0	28.9	21.2	19.1	21.5	23.9	30.1	23.7
	PFENetZ. et al. (2020)	34.3	33.0	32.3	30.1	32.4	38.5	38.6	38.2	34.3	27.4
	ProtRelOkazawa (2022)	42.9	50.6	46.8	47.4	46.9	50.7	58.3	**52.8**	51.3	53.3
	SSPFan et al. (2022)	39.1	45.1	42.7	41.2	42.0	47.4	54.5	50.4	49.6	50.2
	DCAMAShi et al. (2022)	41.5	46.2	45.2	41.3	43.5	48.0	58.0	54.3	47.1	51.9
	Baseline-HSNet	37.2	44.1	42.4	41.3	41.2	45.9	53.0	51.8	47.1	49.5
	Ours	**45.48**	**56.47**	**51.74**	**49.84**	**50.88**	48.87	**61.10**	55.58	**54.03**	**54.90**

modifications to the class partitioning, these two well-known benchmark datasets for semantic segmentation can be utilized to perform few-shot semantic segmentation. Both datasets are partitioned into four folds, with three-quarters of the classes serving as training data (base/seen classes) and the remaining classes serving as validation data (novel/unseen classes). For validation purposes, 1000 episodes of support and query images are sampled from the validation set during the inference phase.

Measures. Results are reported using mean intersection-over-union (mIoU) on individual folds, as well as the average of mIoUs across all folds for both datasets.

Results. Table 3.2 shows the quantitative evaluation on the four folds of the Pascal-5^i dataset. All measures are reported according to their original publications. The highest values are displayed in **bold** and the second-highest appear underlined. We use Min et al. Min et al. (2021) as a baseline since it has similar architecture to ours and similar number of trainable parameters. Following the few-shot semantic segmentation literature, we focus our comparisons on methods reporting on the two backbones Resnet-50 and Resnet-101. For the purposes of a fair comparison, values that differ less than 0.35% are considered equivalent. Except for Shi et al. Shi et al. (2022), which is included

Table 3.4: **Qualitative results.** The first and second columns show the support and query images, respectively, overlaid with the ground truth in red. The remaining columns show the predictions overlaid with a red.



in the comparisons, recent transformer-based methods on few-shot semantic segmentation cannot be integrated with convolutional backbones and are thus excluded because the performance boost is attributed to the change in architecture rather than the methodology G. Sun, Liu, Liang, and Van Gool (2021); J. Zhang et al. (2022). A clear example of this is Shi et al. (2022) which without the Swin-B tranformer backbone the authors report a drop by 5% in the mIoU. Our argument is also supported by the experiments reported by the authors in Shi et al. (2022) where they demonstrate that our baseline Min et al. (2021) when used with a Swin-B backbone gains an average boost of about 6% on mIoU for COCO-20^i 1-shot and 5-shot tasks.

With a Resnet-101 backbone, our method is state-of-the-art for both 1-shot and 5-shot. It exceeds the baseline Min et al. (2021) by 4.21% on 1-shot and 2.72% on 5-shot task. Additionally,

its margins for the 1-shot and 5-shot are 2.51% and 1.12%, compared to the second-best performing methods. With a Resnet-50 backbone, we achieve results comparable to other methods with similar number of trainable parameters. The most recent work of Sun et al. Y. Sun et al. (2022), which achieves state-of-the-art with a Resnet-50, significantly increases the memory requirements because, according to the authors, it requires $128G$ for a batch 8 ($16G$ for one image), which is significantly higher than any other few-shot semantic segmentation technique.

The quantitative evaluation of the four folds of the COCO-20^i data set is displayed in the Table 3.3. We achieve state-of-the-art for COCO-20^i with Resnet-101 backbone for both 1-shot and 5-shot. It exceeds the baseline Min et al. (2021) by 9.68% on 1-shot and by 5.40% on 5-shot. In addition, it has a margin of 3.98% and 1.6% over the second-best performing strategy. Using a Resnet-50 backbone, we achieve second-best performance by a margin of 0.35% on 1-shot and 1.12% compared to the significantly more memory-intensive method of Sun et al.

Table 3.4 displays qualitative comparisons using a Resnet-50 backbone. The first and second columns represent the support and query images, while the remaining columns represent the results of SSP Fan et al. (2022), HSNET Min et al. (2021), DCAMA Shi et al. (2022), BAM Lang et al. (2022), and ours (last column). As can be seen, our method can successfully handle challenging cases in which the object in the support differs visually from the object in the query and the visual similarity between the foreground and background is high, as in the second and fourth rows.

3.4.3 Ablations

Our method results in a substantial performance increase. We demonstrate this by applying it to the classification-based method of Min et al. Min et al. (2021). In the subsequent experiments, we use this as a baseline and conduct 32 experiments consisting of a baseline with a classification backbone (with B_{cls}), a baseline with a segmentation backbone (with B_{sem}), a baseline with dual backbones (with B_{cls} + B_{sem}), and a two-pass dual backbone baseline (two-pass with B_{cls} + B_{sem}). For each ablation, we use Resnet-50 and Resnet-101 backbones, and conduct experiments on all folds of Pascal-5^i for 1-shot and 5-shot. The models are trained for 200 epochs with batch of 12 and Adam optimizer with an initial learning rate of $1e-3$.

We begin with an experiment in which the classification backbone used by the baseline Min et

Table 3.5: **Ablations on all components of our method**. Experiments reported for Pascal-5^i with Resnet-50 and Resnet-101 backbones. The highest values are displayed in **bold** and average mIoU is Highlighted.

		1-shot					5-shot				
Backbone	Method	f0	f1	f2	f3	mIoU	f0	f1	f2	f3	mIoU
Resnet-50	Baseline - HSNET	62.80	70.09	60.16	58.98	63.01	69.12	73.67	66.21	65.44	68.61
	with B_{sem}	60.84	68.27	59.70	59.15	61.99	66.85	72.91	66.55	66.02	68.09
	with B_{cls} + B_{sem}	64.87	71.79	64.12	60.97	65.44	70.11	74.48	67.51	67.09	69.79
	two-pass with B_{cls} + B_{sem}	**66.26**	**73.76**	**63.22**	**63.37**	**66.65**	**70.55**	**75.22**	**67.05**	**68.02**	**70.21**
Resnet-101	Baseline - HSNET	66.41	71.51	62.30	61.96	65.55	71.45	75.07	67.10	67.60	70.30
	with B_{sem}	65.70	72.72	64.01	62.21	66.16	71.16	75.94	69.41	68.03	71.14
	with B_{cls} + B_{sem}	68.47	74.02	64.74	63.59	67.71	72.19	76.85	69.76	69.21	72.00
	two-pass with B_{cls} + B_{sem}	**69.22**	**74.49**	**67.20**	**66.81**	**69.43**	**73.28**	**77.01**	**69.94**	**69.64**	**72.46**

al. (2021) is replaced with a semantic segmentation network in order to gain a better understanding of the impact that the type of the backbone can have on the performance. The first (Baseline) and second (with B_{sem}) rows of each table cell display the results for the 1-shot and 5-shot Pascal-5^i tasks, respectively. Using a classification backbone for Resnet-50 is preferable to using a semantic segmentation backbone. The opposite is true for Resnet-101, and this is supported by the outcomes of both 1-shot and 5-shot tasks. As shown in the third row (with B_{cls} + B_{sem}), it is evident that using both types of backbone improves performance, which is supported by the results on both tasks. As explained in the introduction, this is due to the fact that the B_{cls} and B_{sem} backbones capture diverse but distinct visual features. The fourth row (two-pass with B_{cls} +B_{sem}) displays the results of applying our method to the baseline which increases performance by 3.64% and 3.88% for the 1-shot task with Resnet-50 and Resnet-101 backbones, respectively, and a performance increase of 1.6% and 2.16% for 5-shot for Resnet-50 and Resnet-101, respectively as shown in Table 3.5.

3.5 Conclusion

In conclusion, we proposed a novel two-pass end-to-end method for few-shot semantic segmentation that addresses three key problems affecting performance. The approach leverages an ensemble of visual features learned from pretrained classification and semantic segmentation networks with the same architecture to capture rich and diverse information at different depths. Additionally, the pretrained semantic segmentation network serves as a base class extractor to reduce false

positives. The first pass addresses intra-class similarity by matching support foreground features to query features, and the second pass leverages intra-object similarity by learning to suppress false positives and propagating query foreground features. Experimental results on benchmark datasets demonstrate significant improvement in performance with minimal trainable parameters. Specifically, using Resnet-101, the proposed method achieves state-of-the-art performance for both 1-shot and 5-shot Pascal-5^i, as well as on 1-shot and 5-shot COCO-20^i.

3.6 Data availability

The datasets generated and/or analysed during the current study are available in the PASCAL VOC http://host.robots.ox.ac.uk/pascal/VOC/ and COCO https://cocodataset.org/ repositories.

3.7 Supplementary Material

In the supplementary material, we present (i) additional experiments and explanations on the discriminative power of classification and semantic segmentation networks, (ii) additional justification on employing transductive meta-learning for learning intra-class and intra-object similarity, (iii) additional ablations on false positive reduction, (iv) additional ablations demonstrating the performance boost when using the Shannon entropy loss term $\mathcal{L}_{Sh}$, and (v) larger tables and figures with additional results.

3.7.1 Maximizing Discriminative Power

Figures 3.4 and 3.5 show the discriminative power calculated using more than $4,000$ episodes from Pascal-5^i, ρ^k of each backbone at layers $k, 1 \leq k \leq |B_{cls}|$ of the frozen pretrained backbones B_{cls} and B_{sem}. Figures 3.4 shows the results with Resnet-50, and Figure 3.5 shows the results with Resnet-101. The discriminative power ρ^k at layer k is measured as the ratio $\rho^k = \frac{\frac{1}{N}\sum_i^N cos(FG_Q^i, P_S)}{\frac{1}{M}\sum_j^M cos(BG_Q^j, P_S)}$ where P_S is the support prototype calculated by averaging all the foreground support features FG_S. The numerator is the average cosine distance of the N foreground query

features $FG_Q^i, 0 \leq i \leq N$ to the foreground support prototype FG_S, and the denominator is the average cosine distance of the M background query features $BG_Q^j, 0 \leq j \leq M$ to FG_S.

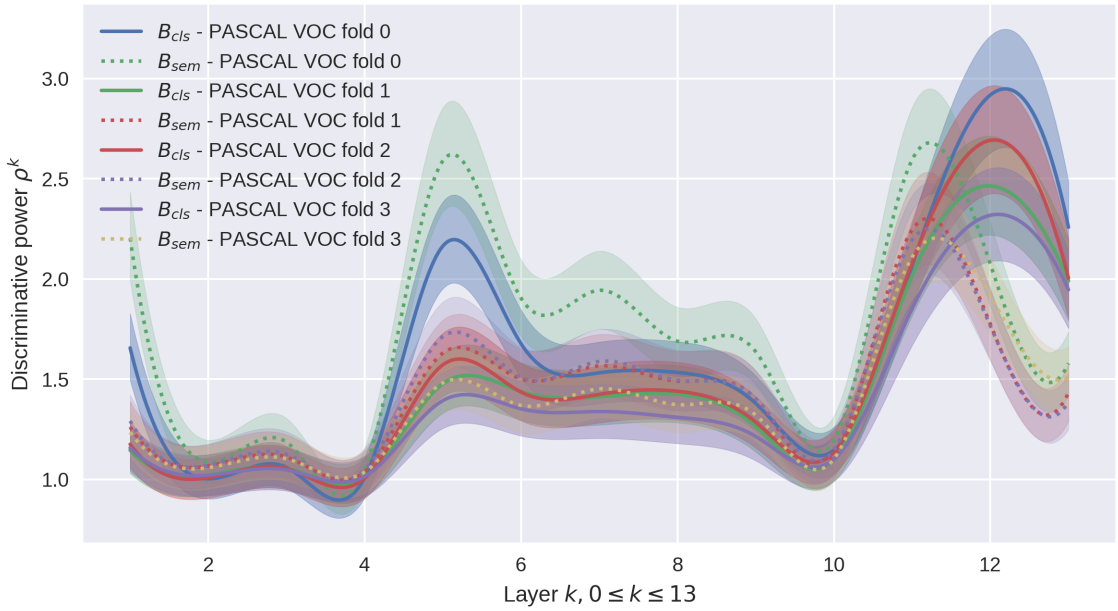


Figure 3.4: **Discriminative power of classification vs semantic segmentation networks.** ——: classification network (Resnet-50), - - -: semantic segmentation network (Resnet-50).

3.7.2 Transductive meta-learning

The object in the support image is frequently not visually similar to that in the query image, leading to under and oversegmentation. Table 3.6, third column (1^{st} pass) depicts the most frequent cases occurring when matching support foreground features to query features, namely undersegmentation (top), and oversegmentation (middle, bottom). Several strategies have been proposed to use this initial query prediction as an additional source of information to improve results in a second step Boudiaf et al. (2021); Fan et al. (2022); K. Wang et al. (2019); C. Zhang et al. (2019). According to the Gestalt principle, the second step can be utilised to refine an undersegmented initial prediction. However, the issue arises when the initial query prediction yields a large number of false positives. To mitigate these cases, SSP method proposed by Fan et al. (2022) presented a two-stage method based on the concept of prototyping for refining the initial query segmentation through the selective propagation of query features in the second step. The selective propagation, which is

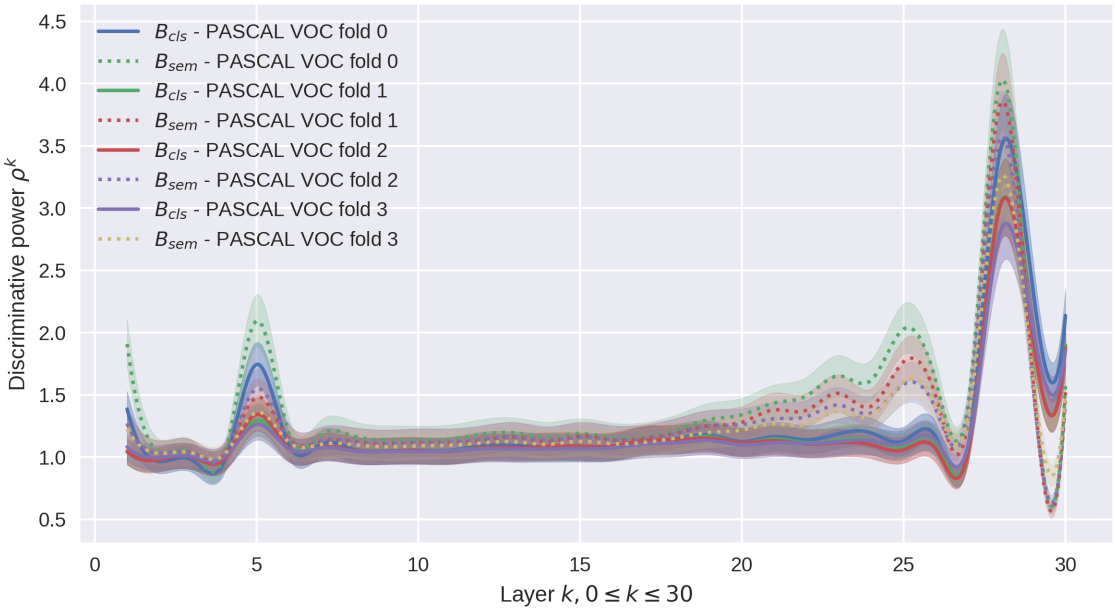


Figure 3.5: **Discriminative power of classification vs semantic segmentation networks.** ——: classification network (Resnet-101), - - -: semantic segmentation network (Resnet-101).

dependent on a user-defined non-adaptive threshold, eliminates gradients and prevents backpropagation throughout the network. In situations where the probability of false positives is greater than the threshold, this not only fails to suppress them but also makes the problem worse by propagating them. Instead of introducing non-differentiable operations like hard-thresholding, as in Fan et al., we address this issue by allowing the network to learn the visual disimilarities between the query foreground features and the false positives in an end-to-end manner. In the first pass, support foreground features are matched to query features, and in the second pass, false positives are suppressed and query foreground features are propagated throughout the query image. The proposed second pass does not introduce new parameters to the network. We use multi-level all-pairs field transforms Teed and Deng (2020) that result in a multiscale hypercorrelation volume Min et al. (2021) to leverage the different levels of visual features learned at each layer of the backbone. Table 3.6, fourth column (2^{nd} pass) demonstrates some instances of our proposed transductive learning method in which the network simultaneously learns to suppress and propagate from initial segmentation. The advantages of our method of self-refinement are summarized below.

(1) We adopted 4D-Conv for our self-refinement module, which outperforms the prototyping

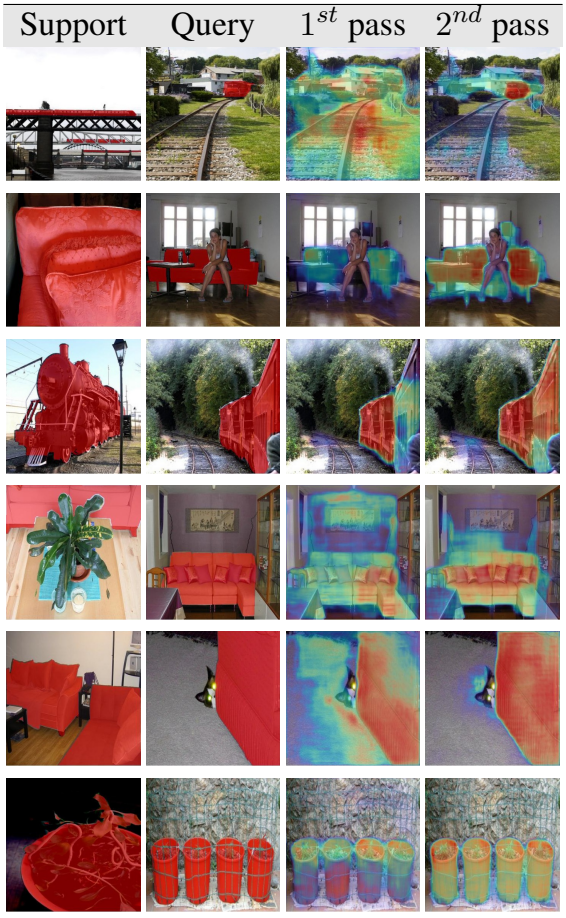
approach of SSP.

(2) Our self-refinement module does not add any additional parameters to the network, whereas the SSP fine-tunes the last two blocks of a ResNet backbone with *1mil parameters*.

(3) Our self-refinement module can operate on top of any backbone, which is another significant advantage over SSP which reshapes embedding space for self-refinement.

(4) SSP employs a non-differentiable method that uses a user-specified hard-threshold. This restricts the ability to add trainable modules after the non-differentiable operation. We do not use non-differentiable operations. Instead, we enable the network to learn end-to-end the visual differences between the query foreground features and false positives.

Table 3.6: **Results from our two-pass method.** 1^{st} pass: intra-class similarity ($S \longrightarrow Q$). 2^{nd} pass: intra-object similarity ($Q \longrightarrow Q$).



3.7.3 Mitigating Propagation of False Positives

The propagation of false positives can be a significant problem in semantic segmentation, particularly when dealing with complex backgrounds or multiple classes that share similar visual features. As noted by Lang et al. Lang et al. (2022), the presence of base classes in the background of the query image can lead to false positive predictions, as the network may incorrectly classify pixels that are not part of the object of interest. To address this issue, they proposed auxiliary layers on top of a base learner that is trained on base classes to predict whether or not each pixel in the output of the meta learner corresponds to a base class. By using this information to selectively mask out base class predictions, they were able to reduce the number of false positives and improve segmentation accuracy. Specifically, Lang et al. proposed an adjustment factor ψ, namely the Frobenius norm of subtraction of Gram matrices, for integrating the output of the base learner and the meta learner.

$$\psi = ||G_s - G_q||_F \tag{17}$$

$$G = ||f_{low}||.||f_{low}^T|| \in R^{C \times C} \tag{18}$$

$$f_{low}^{S,Q} \in R^{C \times H \times W} \tag{19}$$

where $||.||_F$ is the Frobenius norm, f_{low} is the low-level features, and C, H, W represent the embedding dimension, height, and width of the low-level feature maps, respectively. However, our experiments have shown that the adjustment factor ψ is consistently ignored for combining information. Since the Gram matrix is a product of normalised matrices, its values are in the range $[0, 1]$ and as a result, the adjustment factor $\psi \in [0, 1]$. Experimentally, we calculated the adjustment factor of thousands of episodes from train and test sets from the entire COCO-20^i and Pascal-5^i datasets, and observed a mean $\mu_\psi < 0.1$ with a variance $\sigma_\psi < .05$. Table 3.7 shows the results of the experiments on all folds of COCO-20^i and Pascal-5^i datasets. Furthermore, multiplying the network's weights, which are already small, by the adjustment factor, results in near-zero weights that are $20\times$ less than the weight assigned to the meta learner segmentation map. Therefore, we can conclude that the network disregards the adjustment factor in Lang et al. (2022) when integrating the outputs of base and meta-learners. Based on this observation, we propose a method for reducing false positives caused by base classes that is both simpler and faster, resulting in a shorter training

time with the same functionality and performance.

Table 3.7: **Adjustment factor.** The adjustment factor ψ used to combine information in the base learner of Lang et al. (2022) has a mean $\mu_\psi < 0.1$ and a variance $\sigma_\psi < .05$. The small adjustment factor, in conjunction with the fact that these factors are multiplied by the weights which also have small values, leads to near-zero weight compared to the weight assigned to the meta learner segmentation map, which is $20\times$ higher. Below we show the range of ψ after 10,000 episodes on Pascal-5^i and 70,000 episodes on COCO-20^i.

Dataset	Measure	f0	f1	f2	f3
Pascal-5^i	Average	0.087	0.083	0.082	0.081
	Min	0.051	0.039	0.047	0.044
	Max	0.181	0.199	0.162	0.169
COCO-20^i	Average	0.077	0.073	0.067	0.087
	Min	0.036	0.039	0.041	0.037
	Max	0.179	0.163	0.174	0.162

3.7.4 Learning intra-class similarity $S \rightarrow Q$: The impact of the Shannon entropy loss term

The 1^{st} pass is supervised by,

$$\mathcal{L}_{combined} = \frac{1}{N}\sum_{n=1}^{N}[CE(BG_Q^1 \oplus FG_Q^1, Q_n^{gt}) - \kappa\mathcal{L}_{Sh}]$$

where $\kappa = 0.1$. The second term of $L_{combined}$ is the transductive loss term given by Shannon entropy $\mathcal{L}_{Sh}$ given by,

$$\mathcal{L}_{Sh} = \frac{1}{H\times W}\sum_{p=1}^{H\times W}(BG_Q^1(p) \oplus FG_Q^1(p))log(BG_Q^1(p) \oplus FG_Q^1(p)) \tag{20}$$

where $p \in H \times W$ is pixel. The Shannon entropy encourages the network to have a polarised initial prediction with a high or low confidence area S. Dhillon et al. (2020), which reduces the number of false positives.

3.7.5 Additional ablation

Table 3.8 shows the results of supervising the first pass with (bottom row) and without (top row) the Shannon entropy loss term $\mathcal{L}_{Sh}$. The experiments employ our method i.e. two-pass with B_{cls} +

B_{sem}, for both 1-shot and 5-shot tasks on all folds of Pascal-5^i with Resnet-101 backbones for B_{cls} and B_{sem}. As it is evident, there is an improvement in the mIoU for each fold as well as the overall mIoU for both tasks.

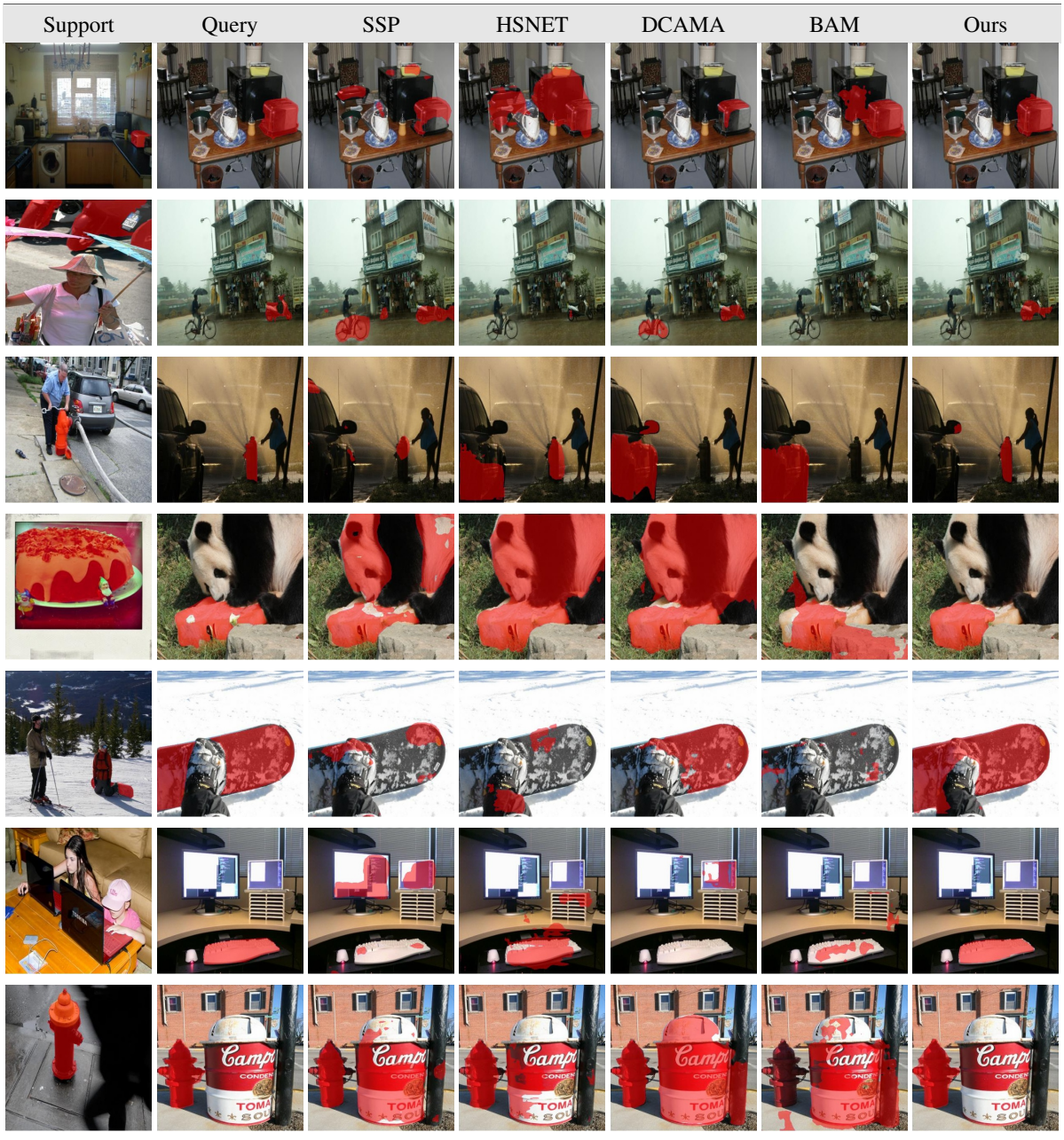
Table 3.8: **Ablation.** Supervising $1^{st}pass$ with (bottom row - w) and without (top row - w/o) the Shannon entropy loss term $\mathcal{L}_{Sh}$.

		1-shot					5-shot				
Backbone	Method	f0	f1	f2	f3	mIoU	f0	f1	f2	f3	mIoU
Resnet-101	w/o $\mathcal{L}_{Sh}$	68.11	74.11	66.05	66.17	68.61	71.08	76.13	69.65	69.12	71.49
	w $\mathcal{L}_{Sh}$	69.22	74.49	67.20	66.81	69.43	73.28	77.01	69.94	69.64	72.46

3.7.6 Additional results

In Table 3.9, we present additional qualitative results, and visual comparisons with the state-of-the-art methods.

Table 3.9: **Qualitative results.** The first and second columns show the support and query images, respectively, overlaid with the ground truth in red. The remaining columns show the predictions overlaid with a red.



Chapter 4

DSV-LFS: Unifying LLM-Driven Semantic Cues with Visual Features for Robust Few-Shot Segmentation

This chapter is a verbatim copy of the paper titled "DSV-LFS: Unifying LLM-Driven Semantic Cues with Visual Features for Robust Few-Shot Segmentation" authored by A. Karimi, C. Poullis, and published in IEEE/CVF Computer Vision and Pattern Recognition, 2025.

Abstract

Few-shot semantic segmentation (FSS) aims to enable models to segment novel/unseen object classes using only a limited number of labeled examples. However, current FSS methods frequently struggle with generalization due to incomplete and biased feature representations, especially when support images do not capture the full appearance variability of the target class. To improve the FSS pipeline, we propose a novel framework that utilizes large language models (LLMs) to adapt general class semantic information to the query image. Furthermore, the framework employs dense pixel-wise matching to identify similarities between query and support images, resulting in enhanced FSS performance. Inspired by reasoning-based segmentation frameworks, our method, named DSV-LFS, introduces an additional token into the LLM vocabulary, allowing a multimodal

LLM to generate a "semantic prompt" from class descriptions. In parallel, a dense matching module identifies visual similarities between the query and support images, generating a "visual prompt". These prompts are then jointly employed to guide the prompt-based decoder for accurate segmentation of the query image. Comprehensive experiments on the benchmark datasets Pascal-5^i and COCO-20^i demonstrate that our framework achieves state-of-the-art performance-by a significant margin-demonstrating superior generalization to novel classes and robustness across diverse scenarios. The source code is available at https://github.com/aminpdik/DSV-LFS

4.1 Introduction

Deep neural networks have shown remarkable success in learning visual features from large labeled datasets Dosovitskiy et al. (2020); K. He et al. (2016); Redmon et al. (2016); Redmon and Farhadi (2016); Ren et al. (2015). However, their ability to generalize to new classes diminishes when only a limited labeled data is available. Few-shot learning Finn et al. (2017); Snell et al. (2017); Vinyals et al. (2016) addresses this limitation by enabling models to learn effectively from a small number of labeled examples, similar to human learning.

In the context of image segmentation Long et al. (2015); Z. Tian et al. (2023), which requires pixel-level annotations, few-shot learning provides a resource-efficient solution. Few-shot semantic segmentation methods focus on predicting detailed masks for novel classes using only a limited number of labeled samples (support images). These methods utilize a range of strategies to effectively leverage the limited labeled samples available for segmentation Boudiaf et al. (2021); Fan et al. (2022); Hong et al. (2021); Lang et al. (2022); W. Liu et al. (2020); Okazawa (2022); Shaban et al. (2017); Y. Sun et al. (2022); K. Wang et al. (2019); Z. et al. (2020); C. Zhang et al. (2019). However, current methods face challenges such as overfitting to the feature distribution of the training data during meta-training, leading to misclassification of seen classes as unseen ones. Additionally, occlusion, deformation, or texture differences between query and support images significantly reduce segmentation accuracy. The root cause for these challenges is the incomplete and appearance-biased feature representation of the novel class learned from the limited data available.

Recent methods address these challenges by leveraging class text descriptions, which provide detailed semantic information to improve segmentation performance F. Liu et al. (2023); Ma et al. (2023); Y. Yang, Chen, Feng, and Huang (2023); Zhu, Chen, Ji, Ye, and Liu (2024). These descriptions help models capture nuanced features of novel classes, enhancing generalization and accuracy even with limited support images. Advances in large language models (LLMs) Brown et al. (2020); Touvron et al. (2023) further enable the efficient encoding of this semantic information, offering a more robust integration of textual and visual cues for improved segmentation.

Large language models (LLMs) have shown great potential in few-shot learning, enhancing performance across tasks in both computer vision and natural language processing Brown et al. (2020, 2023); X. Chen, Sun, Yan, et al. (2023); Kojima, Gu, Reid, Matsuo, and Iwasawa (2022); H. Liu, Jiang, Li, et al. (2023); Touvron et al. (2023). In few-shot segmentation (FSS), integrating LLMs to encode textual information has proven effective in addressing key limitations. While earlier FSS methods used language models for auxiliary tasks such as feature extraction S. He, Ding, and Jiang (2023); F. Liu et al. (2023); Xu, Zhao, Lin, and Long (2023) or generating attribute prompts Ma et al. (2023), recent work Zhu et al. (2024) presents the first direct application of LLMs to FSS, achieving notable improvements in segmentation accuracy. Zhu et al. (2024) engineered prompts leveraging both the support set and class description to guide the LLM in performing segmentation on the query image. However, this method has several key limitations, such as requiring multi-stage training and remaining text-centric, with segmentation results generated as textual descriptions that must be then post-processed to produce a segmentation mask. Despite Zhu et al. (2024) having shown significant improvements through the direct application of LLMs in few-shot settings, an important challenge remains: developing a single-stage, end-to-end pipeline that leverages text-based LLMs to efficiently integrate support images and class descriptions for direct query image segmentation.

To generate segmentation directly on the query image, we adapted a prompt-based decoder Kirillov et al. (2023) to harness its ability to integrate image features with user-provided prompts, facilitating the generation of accurate segmentation masks. These prompts guide the decoder in localizing the region of interest within the query image. Building on recent FSS methods, we efficiently utilize both the support set and class descriptions to guide the decoder.

However, generating prompts from class descriptions presents a significant challenge. While

class descriptions provide consistent, general visual information about the object class, the current query image may lack some of these characteristics due to variations in appearance caused by occlusions, lighting conditions, or partial visibility. As a result, directly encoding class descriptions and incorporating them into the FSS pipeline is not efficient. To overcome this challenge, and inspired by the *reasoning segmentation* framework Lai et al. (2024), we introduce an additional token, $< SEM_{prompt} >$, into the LLM vocabulary, which signifies a request for segmentation based on semantic information. We further design a class semantic encoder module based on a multimodal LLM H. Liu, Li, Li, and Lee (2024); H. Liu, Li, Wu, and Lee (2023), which takes both the query image and general class description to generate query-specific semantic information, referred to as semantic prompt. To further enhance performance, our method incorporates a dense matching module that encodes the similarity between query and support images, producing a visual prompt. This visual prompt complements the semantic prompt by providing fine-grained spatial correspondence, enabling the decoder to better align the class-specific features with the query image. By combining these two forms of guidance, the adapted prompt-based decoder effectively mitigates the limitations of traditional FSS pipelines, delivering superior segmentation accuracy in challenging scenarios and achieving state-of-the-art results by a significant margin.

Our main contributions are as follows:

- To the best of our knowledge, this is the first work, that combines large language models (LLMs) fine-tuned for reasoning segmentation Lai et al. (2024), with foundation semantic segmentation models to directly segment in the context of few-shot semantic segmentation.
- We propose a novel single-stage, end-to-end architecture that seamlessly integrates multimodal semantic features from large language models with visual features derived from pixel-level correspondence, resulting in substantial improvements in segmentation accuracy and robustness.
- We conduct comprehensive experiments across multiple benchmark datasets, demonstrating that our method achieves state-of-the-art, outperforming existing methods by a significant margin.

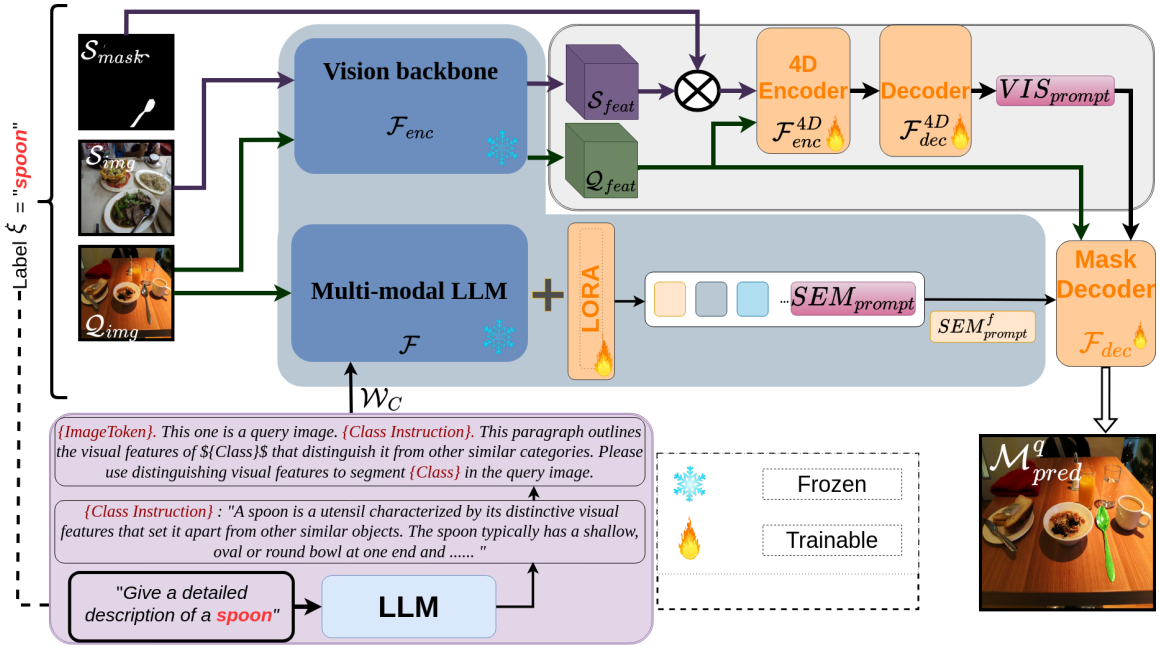


Figure 4.1: **Technical Overview.** The large language model (LLM) first generates a class description W_C based on an input prompt, which consists of a simple question regarding the visual features that distinctly define the class C with label ξ. The $\{ImageToken\}$ in W_C serves as a default token assigned to the query image, and $\{Class\}$ refers to the class label ξ. This class description, along with the query image, is then fed into a multi-modal LLM $(\mathcal{F})$ to produce a class-specific semantic prompt SEM^f_{prompt}. In parallel, a dense matching module $\mathcal{F}^{4D}_{enc}$, $\mathcal{F}^{4D}_{dec}$, generates a class-specific visual prompt VIS_{prompt} by using the support and query feature maps obtained from the vision backbone encoder $\mathcal{F}_{enc}$. Finally, these two prompts, together with the query feature maps, are passed to the prompt-based decoder $\mathcal{F}_{dec}$ to produce the final segmentation.

4.2 Related Work

Few-Shot Segmentation. Classical semantic segmentation methods often rely on a large number of training samples to achieve accurate results. However, to reduce this dependency, few-shot segmentation has emerged as a promising alternative, enabling the segmentation of query images with only a few annotated support images. This approach has received significant attention in recent years. Most FSS methods rely on matching query and support images using prototype-based W. Liu et al. (2020); Siam, Oreshkin, and Jagersand (2019) or pixel-wise methods H. Chen, Dong, Lu, Yu, and Han (2022); Hong et al. (2021); Min et al. (2021); Shi et al. (2022). In prototype-based methods, prototypes are extracted from support images and used for segmentation using either parametric or non-parametric approaches. Non-parametric Dong and Xing (2018); G. Li et al. (2021);

K. Wang et al. (2019); L. Yang, Zhuo, Qi, Shi, and Gao (2021); X. Zhang, Wei, Yang, and Huang (2020) methods classify pixels in the query image based on their similarity to the support prototype, while parametric methods Lang et al. (2022); W. Liu et al. (2020); Wu, Shi, Lin, and Cai (2021); Xie, Liu, Xiong, and Shao (2021); C. Zhang et al. (2019) use learnable parameters to segment the query image using aggregated features of the support and query images. Prototype-based methods in FSS can result in a significant loss of information, as pixel-level features are averaged into a single prototype. To address this limitation, recent work has focused on learning pixel-wise correlations between query and support features, offering a more detailed and accurate approach to segmentation. However, these methods still rely on the limited information available in support images, which may not be sufficient for robust segmentation. To overcome these challenges, several techniques have been developed to incorporate additional information into the few-shot segmentation process. For example, Lang et al. (2022) utilizes a fully supervised, pre-trained network on seen classes to generate a prior map, reducing confusion between seen and unseen classes. Y. Yang et al. (2023) enhances segmentation by incorporating class semantic information, encoding class names using word2vec models, and integrating this information into the query-support matching pipeline for better segmentation outcomes. Zhu et al. (2024), directly employs LLMs for segmentation by introducing detailed task instructions and fine-grained in-context guidance, simulating human cognition to enhance LLMs ability to generate accurate segmentation by providing refined multimodal references.

Large Language Models. LLMs have driven substantial advancements in machine learning, significantly altering the ways in which various tasks are addressed and solved Brown et al. (2020); Touvron et al. (2023); Zhao, Zhou, Li, et al. (2023). These models excel at generating text that closely emulates human language, and they demonstrate exceptional versatility across a wide range of tasks, including transfer learning, few-shot learning, and zero-shot learning Brown et al. (2020, 2023); Kojima et al. (2022); Team (2023); Wei et al. (2022); Zhao, Zhou, et al. (2023). Recently, a number of multi-modal LLM models have been proposed for tasks involving image reasoning, which require a deep integration of visual and textual information to enhance comprehension and interpretation of images. For instance, models like those in H. Liu et al. (2024); H. Liu, Li, et al. (2023) combine the strengths of large language models with advanced visual processing to manage

tasks that require understanding both text and images, such as visual question answering and image captioning. Similarly, LISA Lai et al. (2024) presents a model that incorporates large language models with image segmentation capabilities. This approach uses a $< SEG >$ token to encode input prompts, and the resulting last-layer embedding is decoded into a segmentation mask through the model decoder, leveraging visual features extracted by a vision backbone like SAM Kirillov et al. (2023). Our method improves the few-shot segmentation process by utilizing multimodal LLMs to adapt detailed object class descriptions dynamically to the query image. This semantic information is seamlessly incorporated into the query and support image matching pipeline to guide a prompt-based decoder, enabling more accurate segmentation.

4.3 Method

4.3.1 Problem Definition

Few-shot segmentation focuses on segmenting a target object in a query image with the help of only a few annotated support images. This approach uses meta-learning, where the model is trained using episodes instead of conventional image batches. Each episode consists of a support set and a query set. In a K-shot setting, the support set $S = \{X_i^s, M_i^s\}_{i=1}^K$ includes K support images X^s and their corresponding masks M^s, while the query set $Q = \{X^q, M^q\}$ comprises a query image X^q and its corresponding mask M^q used for the loss calculation during training.

In the standard few-shot segmentation setting, the support set provides annotated examples for a target class $C \in C_{train} \cup C_{test}$, inherently including the class label through the segmentation masks. Our method leverages this class label C to generate a class description W_C using ChatGPT, which enriches the semantic understanding of the target class without introducing additional information beyond what is available in traditional FSS tasks. The episodes are drawn from the training dataset $D_{train} = \{(S_i, Q_i, W_C)\}_{i=1}^{N_{train}}$ for the meta-training phase and from the testing dataset $D_{test} = \{(S_i, Q_i, W_C)\}_{i=1}^{N_{test}}$ for meta-testing. D_{train} has C_{train} classes, and D_{test} has C_{test} classes, with no overlap between the two, i.e., $C_{train} \cap C_{test} = \emptyset$. The goal is to train the model on D_{train} and then test it on unseen classes in D_{test}, leveraging the learn-to-learn paradigm.

4.3.2 Overview

The proposed few-shot semantic segmentation network comprises three core modules, as depicted in Figure 4.1. First, the Class Semantic Encoder $\mathcal{F}$ employs a multimodal LLM to adapt the general class description to the query image, generating the $< SEM_{prompt} >$. Second, the Dense Matching Module identifies visual correspondences between the query and support images, producing the VIS_{prompt}. Finally, the prompt-based Decoder Module $\mathcal{F}_{dec}$ combines these two complementary prompts with query image features extracted by a visual encoder $\mathcal{F}_{enc}$ to accurately segment the target object.

4.3.3 Class Description Generation

To generate the class descriptions, we query ChatGPT 4.0 with a specific prompt for each class with label ξ as follows: Please provide a detailed description of the visual characteristics that uniquely identify the $< classname >$ object class, distinguishing it from other similar object categories. Focus solely on the distinguishing visual features in a comprehensive paragraph. This prompt ensures that the descriptions are tailored to highlight the unique visual features of each class. We collect the responses from ChatGPT for all classes. As an example, the following is a class description generated for "spoon": A spoon is a utensil characterized by its distinctive visual features that set it apart from other similar objects. The spoon typically has a shallow, oval or round bowl at one end, designed to hold and scoop liquids or semi-solids. The handle of a spoon is usually elongated, straight, or slightly curved, and tapers towards the bowl, allowing for a comfortable grip. They are commonly made from reflective materials like stainless steel, which give them a shiny appearance, but can also be found in other materials such as wood or plastic, each presenting a different texture and finish. Unlike forks, spoons lack tines, and unlike knives, they do not have a blade or sharp edges, making their overall form smooth and rounded.

We experimented with various prompt formulations to mitigate sensitivity in LLM outputs. Our final prompt structure described above emphasizes distinctive visual features, and we found this approach to consistently generate informative and reliable class descriptions.

Incorporating the generated class descriptions, and following the approach of H. Liu et al.

(2024); H. Liu, Li, et al. (2023), we construct input prompt for class encoder module ($\mathcal{W}_C$ in Figure 4.1) as: $ImageToken$. This one is a query image. $ClassInstruction$. This paragraph outlines the visual features of $Class$ that distinguish it from other similar categories. Please use distinguishing visual features to segment $Class$ in the query image. The $ImageToken$ serves as a default token assigned to the query image within the input prompt. This token is subsequently replaced by the output features of the query image generated by CLIP Radford et al. (2021). The $ClassInstruction$ corresponds to the specific description generated for the target class, while the $Class$ represents the name of the object category being segmented. The expected output follows this format: Sure, the segmentation result is $< SEM_{prompt} >$, where $< SEM_{prompt} >$ is a token added to the LLM vocabulary. This token enables the LLM to adapt the general class description to the visual features of the target object, providing query image-specific semantic information.

4.3.4 Class Semantic Encoder Module

Class descriptions offer general visual information about an object class, but query images often vary due to factors such as occlusions, lighting changes, or partial visibility. As a result, directly incorporating class descriptions into the FSS pipeline is not efficient. To address this, we propose a class semantic encoder module that adapts class descriptions to the query image, generating a context-aware semantic prompt that captures the specific characteristics of the target class within the query image. To enable this adaptation, we introduce the $< SEM_{prompt} >$ token into the LLM vocabulary, which requests detailed, query image-specific semantic information. Following the LLaVA architecture, features from the query image $Q^{\mathcal{C}}_{img}$ containing the target object class $\mathcal{C}$, extracted by a visual encoder, i.e., CLIP Radford et al. (2021), along with the prepared description $W_{\mathcal{C}}$ for the target object class, are fed into the LLM, which generates a text response.

$$y_{text} = \mathcal{F}(Q^{C}_{img}, W_{\mathcal{C}}) \tag{21}$$

When instructed to generate a segmentation aligned with the target class description, the output y_{text} includes the $< SEM_{prompt} >$ token, which encodes semantic information specific to the target class in the context of the query image. Next, we extract the LLM last-layer embedding, h_{sem}

, corresponding to the $< SEM_{prompt} >$ token, and apply an MLP projection layer to obtain the final semantic prompt SEM^{f}_{prompt}.

$$SEM^{f}_{prompt} = MLP_{proj}(h_{sem}) \tag{22}$$

4.3.5 Dense Matching Module

This module generates a visual prompt VIS_{prompt} that encodes the similarity between the target object in the support set and the query image. To achieve this, we leverage dense matching between annotated support images and the query image, which has been shown to outperform prototype-based matching by capturing fine-grained details Min et al. (2021). We extract a diverse range of features from various L depths of a vision transformer, i.e., the SAM Kirillov et al. (2023), $\{(f^l_S, f^l_Q)\}^L_{l=1}$ forming a set of 4D hypercorrelation tensors, $HPV_l \in \mathbb{R}^{H_p \times W_p \times H_p \times W_p}$. These 4D hypercorrelations are then stacked together.

$$HPV_{l_{(X_q, X_s)}} = \text{ReLU}\left(\frac{f^l_Q \cdot f^l_S}{\|f^l_Q\| \|f^l_S\|}\right) \tag{23}$$

$$\textbf{HPV} = \text{Concat}\Big(HPV_1, HPV_2, \ldots, HPV_L\Big) \tag{24}$$

$$\textbf{HPV} \in \mathbb{R}^{L \times H_p \times W_p \times H_p \times W_p} \tag{25}$$

Using efficient center-pivot 4D convolutions ($\mathcal{F}^{4D}_{enc}$ in Figure 4.1), the method combines high-level semantic and low-level geometric cues from hypercorrelations to encode matching of support and query images. The encoder output is then passed to a decoder module ($\mathcal{F}^{4D}_{dec}$ in Figure 4.1), which generates a visual prompt, VIS_{prompt}, to guide the segmentation process.

$$\mathcal{H}^{4D} = \mathcal{F}^{4D}_{enc}(\textbf{HPV}) \tag{26}$$

$$VIS_{prompt} = \mathcal{F}^{4D}_{dec}(\mathcal{H}^{4D}) \tag{27}$$

4.3.6 Mask Decoder Module

In our method, we employ two types of prompts, semantic (SEM^f_{prompt}) and visual (VIS_{prompt}) to guide the prompt-based decoder ($\mathcal{F}_{dec}$ in Figure 4.1) for accurate segmentation of the query image. The semantic prompt, produced by the Class Semantic Encoder, captures detailed attributes of the target object, such as shape, texture, and distinctive features, providing contextual information for segmentation. The visual prompt, generated by the Dense Matching Module, is derived from pixel-wise matching between the query image and annotated support images, identifying the target regions in the query image. These prompts, combined with query image features Q_{feat}, extracted by the encoder $\mathcal{F}_{enc}$, are integrated into the prompt-based decoder $\mathcal{F}_{dec}$ Kirillov et al. (2023) to directly generate segmentation on the query image.

$$M^q_{pred} = \mathcal{F}_{dec}(Q_{feat}, VIS_{prompt}, SEM^f_{prompt}) \tag{28}$$

The decoder employs multi-head attention and transformer blocks to fuse the semantic and visual information, refining mask proposals through learnable queries. This hierarchical approach merges the high-level context from the semantic prompt with spatial details from the visual prompt, producing detailed segmentation masks. This integration enables the decoder $\mathcal{F}_{dec}$ to deliver segmentation results that are precise and contextually coherent.

4.3.7 Training loss

Our model is trained end-to-end using two main loss functions adapted from Lai et al. (2024): the text generation loss $\mathcal{L}_{\text{text}}$ and the segmentation mask loss $\mathcal{L}_{\text{mask}}$. The overall loss function $\mathcal{L}$ is defined as the weighted sum of these two losses:

$$\mathcal{L} = \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}, \tag{29}$$

where λ_{text} and λ_{mask} are the weights assigned to the text and mask losses, respectively. The text generation loss $\mathcal{L}_{\text{text}}$ is formulated as the auto-regressive cross-entropy loss, while the segmentation mask loss $\mathcal{L}_{\text{mask}}$ combines per-pixel binary cross-entropy (BCE) loss and Dice loss, weighted by

λ_{BCE} and λ_{Dice}.

Given the ground truth text labels $\hat{\mathbf{y}}_{\text{text}}$ and query mask $\mathbf{M}^q$ and predicted mask for query image $\mathbf{M}^q_{pred}$, the specific loss functions are:

$$\mathcal{L}_{\text{text}} = \text{CE}(\hat{\mathbf{y}}_{\text{text}}, \mathbf{y}_{\text{text}}) \tag{30}$$

$$\mathcal{L}_{\text{mask}} = \lambda_{\text{BCE}}\text{BCE}(\mathbf{M}^q_{pred}, \mathbf{M}^q) + \tag{31}$$

$$\lambda_{\text{Dice}}\text{Dice}(\mathbf{M}^q_{pred}, \mathbf{M}^q) \tag{32}$$

Our method integrates these losses to extend the capabilities of multimodal large language models (LLMs), enabling them to handle both text generation and fine-grained segmentation tasks.

4.3.8 Extending to K-shot setting

For the K-shot scenario, we adopt the strategy proposed in Min et al. (2021). With K support image-mask pairs and a query image, the model makes K separate forward passes, resulting in K predicted segmentation masks. To determine the final segmentation, a voting mechanism is applied at each pixel, where the sum of the K predictions is normalized by the maximum possible votes. Pixels are then classified as foreground if their normalized score exceeds a certain threshold, allowing for a more robust segmentation decision based on multiple support examples.

4.4 Experiments

4.4.1 Experimental Settings

Benchmark datasets. Following previous works in FSS, we evaluate the proposed method on two few-shot segmentation benchmark datasets: Pascal-5^i and COCO-20^i, derived from PASCAL VOC 2012 and MS-COCO, respectively. Each dataset is divided into four folds, with three-quarters of the classes designated for training (base/seen classes) and the remainder for testing (novel/unseen classes). During the inference phase, 1000 episodes of support and query images are randomly

sampled from the test set to evaluate the model performance.

Evaluation measures. To evaluate the proposed method, we adopt mean intersection-over-union (mIoU), consistent with previous studies. To ensure robust and reliable results, we conduct five trials for each experiment using different random seeds. The final performance metric is obtained by averaging the results from all five trials, providing a comprehensive assessment of the method effectiveness for few-shot segmentation tasks.

Implementation details. The proposed network combines the pre-trained multi-modal language model LLaVA (llava-v1.5-7b) H. Liu et al. (2024) with the Segment Anything network Kirillov et al. (2023). The network introduces a 4D dense matching module, which utilizes center-pivot 4D convolutions Min et al. (2021) followed by a convolutional decoder module. The mask decoder is derived from the Segment Anything mask decoder Kirillov et al. (2023). To generate class descriptions, a custom Python web scraping tool is used to query ChatGPT 4.0, producing detailed descriptions for all classes in the benchmark datasets.

Training. One advantage of the proposed model is that it functions as an end-to-end model trained in a single stage. To efficiently fine-tune the multi-modal LLM, we employ LoRA E. J. Hu et al. (2021) while keeping the vision backbones frozen. Meanwhile, the 4D encoder/decoder and mask decoder are fine-tuned, and the LLM token embeddings, language model head (LM head), and projection layer are set as trainable parameters. The batch size is set to 2 per GPU, and the model is trained for 10 epochs using the AdamW optimizer with the cosine annealing scheduler and an initial learning rate of $3e-4$. The loss weights are set to 1, 2, and 0.5 for $\lambda_{\text{text}}, \lambda_{\text{BCE}}, \lambda_{\text{Dice}}$, respectively. To ensure a fair comparison with other methods, no data augmentation is used during training. Two NVIDIA A100 GPUs are employed for training.

Fairness in Comparisons. While our method incorporates detailed class descriptions generated from the class label, we maintain consistency with the standard FSS setting where the class label is inherently available through the support set annotations. Previous works have also leveraged class semantics, such as class names or word embeddings Xian, Choudhury, He, Schiele, and Akata (2019); Y. Yang et al. (2023) and more recently language guidance J. Wang, Liu, Zhou, and Wang (2024); Zhu et al. (2024), to enhance segmentation performance in FSS.

4.4.2 Comparison with State-of-the-Art

Table 4.1: **Performance Comparisons.** We evaluate our method by comparing the mean intersection-over-union (mIoU) on the PASCAL-5^i and COCO-20^i datasets against other state-of-the-art methods. To ensure the robustness and reliability of the results, we perform each experiment five times using different random seeds and report the average mIoU scores for both 1-shot and 5-shot settings. The highest values are indicated in **bold**, the second-highest are underlined, and the average mIoU is highlighted.

Dataset	Method	Conference	1-shot					5-shot				
			Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
PASCAL-5^i	NTRENet Y. Liu, Liu, Cao, et al. (2022)	CVPR'22	65.4	72.3	59.4	59.8	63.2	66.2	72.8	61.7	62.2	65.7
	BAM Lang et al. (2022)	CVPR'22	69.0	73.6	67.5	61.1	67.8	70.6	75.1	70.8	67.2	70.9
	AAFormer Y. Wang, Sun, Zhang, and Zhang (2022)	ECCV'22	69.1	73.3	59.2	65.2	66.7	72.5	74.7	62.0	61.3	67.6
	SSP Fan et al. (2022)	ECCV'22	60.5	67.8	56.1	61.4	61.5	67.5	72.7	75.2	62.1	69.3
	IPMT Y. Liu, Liu, Yao, and Han (2022)	NeurIPS'22	72.8	73.7	59.2	61.6	66.8	73.1	74.7	61.6	63.4	68.2
	ABCNet Y. Wang, Sun, and Zhang (2023)	CVPR'23	68.8	73.4	59.6	65.0	66.5	71.7	74.2	74.8	67.0	69.6
	HDMNet Peng et al. (2023)	CVPR'23	71.0	75.4	62.1	69.4	69.7	71.3	76.2	71.3	68.5	71.8
	MIANet Y. Yang et al. (2023)	CVPR'23	68.5	75.8	64.5	68.7	69.4	70.2	77.4	70.0	68.8	71.7
	MSI Moon et al. (2023)	ICCV'23	71.0	72.5	63.8	65.9	68.3	73.0	74.2	70.5	66.6	71.1
	SCCAN Xu et al. (2023)	ICCV'23	68.3	72.5	66.8	58.9	66.6	72.3	74.1	69.1	65.6	70.3
	LLaFS Zhu et al. (2024)	CVPR'24	**74.2**	78.8	**72.3**	68.5	73.5	**75.9**	80.1	**75.8**	70.7	**75.6**
	DSV-LFS		71.67	**81.97**	71.17	**75.04**	**74.96**	72.03	**82.01**	71.32	**75.51**	75.21
COCO-20^i	NTRENet Y. Liu, Liu, Cao, et al. (2022)	CVPR'22	36.8	42.6	39.7	39.3	38.2	38.2	44.1	40.4	38.4	40.3
	BAM Lang et al. (2022)	CVPR'22	43.4	50.6	47.5	43.6	46.3	49.3	54.2	51.6	49.9	51.2
	SSP Fan et al. (2022)	ECCV'22	35.5	39.6	37.9	36.7	37.4	40.6	47.0	45.1	43.9	44.1
	AAFormer Y. Wang et al. (2022)	ECCV'22	39.8	44.6	41.1	41.6	41.8	42.9	50.1	45.5	49.6	49.6
	MM-Former G. Zhang et al. (2022)	NeurIPS'22	40.5	47.7	45.2	43.4	44.2	40.4	47.4	50.0	48.8	46.6
	IPMT Y. Liu, Liu, Yao, and Han (2022)	NeurIPS'22	41.4	45.2	45.6	40.4	43.2	43.3	47.5	43.8	42.5	44.3
	ABCNet Y. Wang et al. (2023)	CVPR'23	42.3	46.2	46.0	44.1	44.7	44.5	51.7	52.2	46.4	49.1
	HDMNet Peng et al. (2023)	CVPR'23	43.8	50.8	50.6	49.4	48.6	50.6	61.6	55.7	56.6	56.1
	MIANet Y. Yang et al. (2023)	CVPR'23	42.5	50.3	47.8	47.4	47.7	45.8	58.2	51.3	51.9	51.7
	MSI Moon et al. (2023)	ICCV'23	42.4	47.4	44.9	44.6	44.8	47.1	53.2	53.4	51.9	51.9
	SCCAN Xu et al. (2023)	ICCV'23	40.4	42.6	41.4	40.7	41.3	47.2	57.2	59.2	52.1	53.9
	LLaFS Zhu et al. (2024)	CVPR'24	47.5	58.8	56.2	53.0	53.9	53.2	63.8	63.1	60.0	60.0
	DSV-LFS		**69.97**	**73.35**	**70.69**	**71.32**	**71.33**	**71.05**	**73.81**	**71.32**	**71.45**	**71.90**

Pascal-5^i: Table 4.1 presents a comparison of the mIoU measure between our method and several recent few-shot segmentation approaches on the Pascal-5^i benchmark. The results demonstrate that while our method significantly surpasses all non-LLM-based approaches, it achieves comparable results with the LLM-based approach of Zhu et al. (2024). We attribute this to the dataset being simpler than other benchmarks, as it contains fewer classes and only one object class per image, resulting in potential saturation. This hypothesis is further supported by the results on other benchmark datasets, which surpass *all* other methods by a significant margin, as we describe below.

COCO-20^i: Table 4.1 presents the performance of our proposed method on the COCO-20^i dataset, which is known for its challenging segmentation tasks due to the presence of multiple

objects and significant intra-class variability. Our method achieves notable improvements over existing approaches, with gains of +17.43% mIoU in the 1-shot setting and +11.9% mIoU in the 5-shot setting.

COCO-20^i $\rightarrow$ **Pascal-**5^i: On the basis of the different distributions of the training dataset and testing dataset, a model trained on one dataset is evaluated on another without any fine-tuning. To demonstrate the effectiveness of our method, we perform experiments in the COCO-20^i $\rightarrow$ Pascal-5^i. We trained our network on all classes of COCO-20^i and evaluate the network on Pascal-5^i without fine-tuning. As shown in Table 4.2, while our network is not specifically designed for cross-domain few-shot segmentation, it still achieves state-of-the-art results, demonstrating a gain of +1% mIoU in the 1-shot setting compared to other cross-domain FSS methods that are explicitly developed for this purpose.

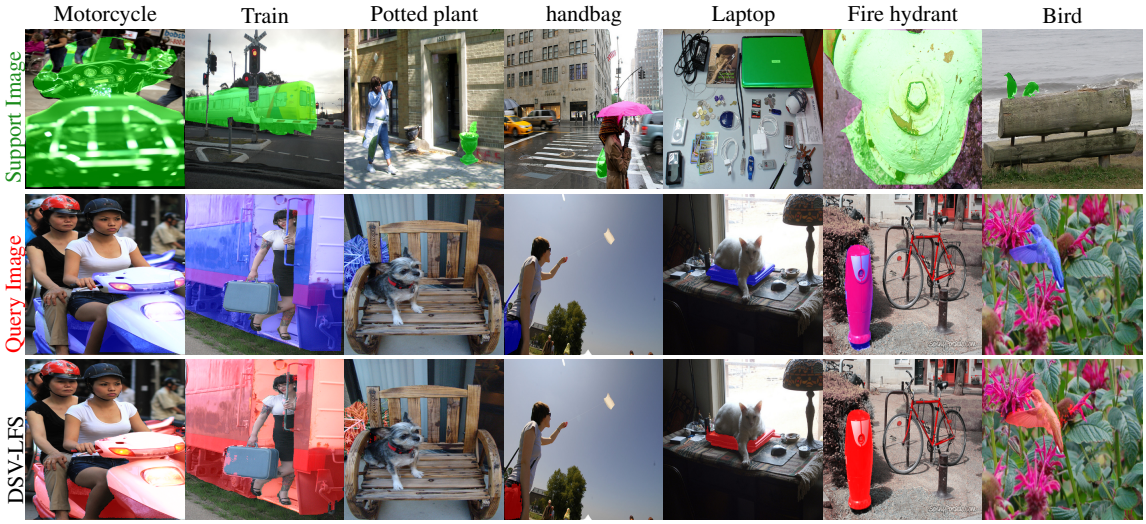


Figure 4.2: **Qualitative results.** Examples of our method's performance on the COCO-20^i dataset. Each column represents an episode, displaying the support image, query image, and predicted segmentation output from top to bottom. The episodes illustrate the model's ability to handle challenges such as the presence of base classes in the query image (e.g., person in motorcycle and train classes) and variations between target objects in support and query images, including scale differences (e.g., handbag), occlusion (e.g., laptop), appearance changes (e.g., potted plant), complex backgrounds (e.g., bird), and deformations (e.g., fire hydrant).

4.4.3 Qualitative Results

We present qualitative results generated by our proposed method to illustrate its effectiveness in overcoming key challenges in few-shot segmentation. Two common issues in this task include: (1) the misclassification of base class objects as novel classes, leading to false positives; and (2) the reliance on a few support images, which often do not capture the full variability of the target class appearance. Figure 4.2 demonstrates our method's robustness against misclassification of base classes as novel classes. The first and second columns highlight how our method accurately distinguishes target classes such as "motorcycle" and "train" from other objects in the query images. In the subsequent columns, we showcase the method's ability to precisely segment target classes despite significant variations, including differences in scale (e.g., handbag), occlusion (e.g., laptop), appearance changes (e.g., potted plant), cluttered backgrounds (e.g., bird), and deformations (e.g., fire hydrant) between the support and query images. These results underline the adaptability of our method to handle complex and diverse visual scenarios, substantially improving segmentation performance in few-shot learning contexts.

Table 4.2: **Performance comparison of our model on the COCO-20^i $\rightarrow$ Pascal-5^i cross-domain setting, without fine-tuning.** Although our method was not explicitly designed for cross-domain FSS, it achieves state-of-the-art results with a +1 mIoU gain in the 1-shot setting. We run each experiment five times with different random seeds and report the average mIoU for the 1-shot setting. The highest values are indicated in **bold**, the second-highest are underlined, and the average mIoU is highlighted.

COCO-20^i $\rightarrow$ Pascal-5^i					
Methods	**Fold-0**	**Fold-1**	**Fold-2**	**Fold-3**	**mIoU**
PFENet(TPMAI) Z. et al. (2020)	43.2	65.1	66.5	69.7	61.1
RePRI(CVPR'21) Boudiaf et al. (2021)	52.2	64.3	64.8	71.6	63.2
VAT(ECCV'22) Hong et al. (2021)	52.1	64.1	67.4	74.2	64.5
VAT-HM(ECCV'22) W. Liu, Zhang, Ding, Hung, and Lin (2022)	48.3	68.6	69.6	79.8	65.6
HSNet(ICCV'21) Min et al. (2021)	47.0	65.2	67.1	77.1	64.1
HSNet-HM(ECCV'22)W. Liu et al. (2022)	46.7	68.6	71.1	79.7	66.5
RTD(CVPR'22) W. Wang et al. (2022)	59.4	70.4	70.5	78.4	69.7
PMNet (WACV'24) H. Chen et al. (2022)	71.0	72.3	66.6	63.8	68.4
IFA(CVPR'24) Nie et al. (2024)	-	-	-	-	79.6
DSV-LFS	**74.86**	**85.23**	**82.23**	**80.37**	**80.67**

4.4.4 Ablations

To assess the effectiveness of various components and design choices in our method, we conducted extensive ablation studies using the 1-shot setting of the COCO-20^i dataset. We chose COCO-20^i for our ablation study because it is a more challenging dataset with multiple objects per image and significant variability in object scales, poses, and contexts. This complexity makes it ideal for evaluating the robustness of each component of our proposed method. Table 4.3 illustrates how each component contributes to the overall model performance.

Table 4.3: **Ablation.** We evaluate segmentation performance using semantic prompts alone vs a combination of semantic & visual prompts and report the mean intersection-over-union (mIoU) on the **COCO**-20^i dataset. The highest values are indicated in **bold**.

Method	**1-shot**				
	Fold-0	**Fold-1**	**Fold-2**	**Fold-3**	**mIoU**
DSV-LFS w/ semantic prompt only	66.99	71.34	67.52	70.14	68.99
DSV-LFS w/ semantic & visual prompt	**69.97**	**73.35**	**70.69**	**71.32**	**71.33**

Effect of semantic prompt

To assess the effectiveness of class descriptions, we conducted an experiment in which the mask decoder was guided exclusively by the semantic prompt generated by the Class Semantic Encoder Module. As shown in Table 4.3, our method achieved a significant improvement of +15 mIoU over the current state-of-the-art FSS methods, even when relying solely on the semantic prompt. These results underscore the value of leveraging semantic knowledge from large language models in few-shot segmentation tasks.

Effect of visual prompt

An additional ablation was conducted to evaluate the effect of incorporating a small number of labeled samples alongside semantic information. In this experiment, we combined both the visual prompt from support images and the semantic prompt generated from class descriptions. The

integration of these two prompts resulted in an approximate 3% improvement as shown in Table 4.3 compared to using the semantic prompt alone. This demonstrates that the visual prompt effectively complements the FSS pipeline by providing important visual cues that enhance the model ability to generalize to novel classes. The synergy between visual and semantic prompts indicates that leveraging both modalities can more effectively capture the diverse features and variations of target objects, thereby improving overall segmentation accuracy.

4.5 Conclusion

This paper presents a novel approach to few-shot semantic segmentation that uniquely integrates LLM-derived semantic prompts with dense visual matching. We introduce a new token, $< SEM_{prompt} >$, into the LLM vocabulary to generate class-specific semantic prompts, which are combined with visual prompts $< VIS_{prompt} >$ obtained through dense visual matching between query and support images. This dual-prompt strategy is inspired by the way the human brain and visual system rapidly learn and recognize new objects by drawing upon a vast repository of prior knowledge while using the visual features of unfamiliar objects. Similarly, our approach combines the broad knowledge-base of LLMs with object-specific visual features from limited samples, resulting in a robust segmentation performance. Our method addresses the limitations of prior work by providing richer contextual information and achieving superior performance on challenging benchmarks by a significant margin. By integrating semantic and visual cues, it addresses FSS challenges and demonstrates LLMs' potential to improve segmentation and guide future research on complex tasks.

4.6 Supplementary Material

In the supplementary material, we provide qualitative examples with detailed class descriptions.

4.6.1 Qualitative Results

We present qualitative results from our proposed method to demonstrate its effectiveness in addressing key challenges in few-shot segmentation. The two primary challenges in this context are: (1) the misclassification of base class objects as novel classes, resulting in false positives, and (2) reliance on a limited set of support images, which often fails to capture the full range of target class variations.

The following examples illustrate challenging episodes that highlight these issues. In the examples, the input episode consists of annotated support and query images alongside the DSV-LFS output. While the support and query images are annotated to specify the object of interest, the annotation on the DSV-LFS output represents the predictions of the proposed method. Additionally, a detailed class description is provided as an input prompt for the multi-modal LLM to generate the semantic prompt.<Qimage>in the class descriptions serves as a default token assigned to the query image within the input class description. This token is subsequently replaced by the output features of the query image.

Input Episode
Motorcycle


Support Image


Query Image


DSV-LFS

Complete class description for mutli-modal LLM for motorcycle object class:

<Qimage>. this one is a query image. A motorcycle is distinctively characterized by its two-wheeled structure, which sets it apart from other vehicles. The wheels are large and typically exposed, with a prominent front wheel that often includes a visible brake disc and caliper, and a rear wheel that may have a broader tire. The frame is compact and streamlined, with a noticeable absence of a roof or any extensive enclosure. The handlebars, which are prominently situated above the front wheel, feature visible controls and mirrors extending outward. The seat is elongated and generally positioned for a straddling rider, often with a noticeable saddle shape. Beneath the seat, the engine is a dominant visual element, with its intricate metallic components like the exhaust pipes and cylinders often exposed. The fuel tank, typically located in front of the seat and above the engine, is a rounded or angular structure with a glossy finish. Additionally, motorcycles have a distinctive set of front and rear lights; the front light is usually a singular, circular or angular headlamp, while the rear includes a smaller brake light. The suspension system, including visible shock absorbers and forks, also adds to its unique visual identity. A motorcycle is distinguished from other similar object categories by several unique visual features. Primarily, it has two large wheels in tandem with a sleek, streamlined frame connecting them. Additionally, motorcycles usually have larger, more prominent headlights and taillights compared to bicycles, often integrated into the design rather than being detachable. The tires on motorcycles are wider and more robust than those on bicycles, designed to handle higher speeds and more significant weight. These visual features collectively distinguish motorcycles from bicycles, mopeds, and scooters. This paragraph outlines the visual features of motorcycle that distinguish it from other similar categories. Please use distinguishing visual features to segment motorcycle in the query image.

Input Episode
Keyboard


Support Image


Query Image


DSV-LFS

Complete class description for mutli-modal LLM for keyboard object class:

<Qimage>. this one is a query image. A keyboard, in its distinctive visual form, is typified by its flat, elongated shape with an array of rectangular keys arranged in neat rows. Each key is typically square or slightly rectangular, often featuring rounded edges for ergonomic comfort during typing. The surface of the keys is uniformly smooth and matte or glossy, contrasting with the often darker or neutral-colored base. These keys are distinctly marked with alphanumeric characters, symbols, and functional indicators, often in contrasting colors such as white or light gray on dark backgrounds, aiding visibility and usability. Additionally, keyboards commonly include functional sections such as arrow keys, function keys (F1-F12), and a dedicated numerical keypad (on larger models), each section visually demarcated or slightly raised for tactile distinction. The overall profile of a keyboard is thin and flat, designed for ergonomic use on desks or tables, typically with a USB cable or wireless connectivity. These visual features collectively distinguish a keyboard from similar objects like calculators or remote controls, which lack the array of keys and alphanumeric layout essential for text input and control in computing environments. This paragraph outlines the visual features of keyboard that distinguish it from other similar categories. Please use distinguishing visual features to segment keyboard in the query image.

Input Episode
Toilet


Support Image


Query Image

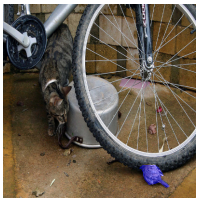

DSV-LFS

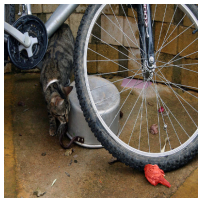
Complete class description for mutli-modal LLM for toilet object class:

<Qimage>. this one is a query image. A toilet is a distinct bathroom fixture characterized by several unique visual features. It typically has a bowl-shaped seat made of porcelain or ceramic, with a rounded or oval opening that slopes inward. The bowl is often connected to a pedestal or base, which is relatively narrow compared to the bowl itself, giving it a recognizable silhouette. Attached to the back of the bowl is a water tank, which is usually rectangular and taller than it is wide, designed to hold flushing water. The toilet seat, often made of plastic, is hinged at the rear and can be lifted or lowered. This seat usually has a lid that matches in material and design. The bowl's interior is smooth and glazed, facilitating easy cleaning and often features a water-filled trap at the bottom, visible when the lid and seat are raised. The flush handle or button is typically located on the side or top of the water tank, which distinguishes it from other fixtures like bidets or urinals that lack such a tank. Overall, the combination of the bowl's shape, the attached water tank, the hinged seat and lid, and the flush mechanism make the toilet visually distinct from similar bathroom objects.This paragraph outlines the visual features of toilet that distinguish it from other similar categories. Please use distinguishing visual features to segment toilet in the query image.

Input Episode
Bird


Support Image


Query Image


DSV-LFS

Complete class description for mutli-modal LLM for bird object class:

<Qimage>. this one is a query image. Birds are characterized by their distinctive features, which set them apart from other similar object categories. Birds possess a unique feather covering, often brightly colored or patterned, providing insulation and aiding in flight. Their beaks vary in shape and size depending on their diet, from sharp, curved beaks in birds of prey to flat, broad ones in filter feeders. They have lightweight, streamlined bodies adapted for flight, with a high degree of symmetry and hollow bones. Their wings, a key identifier, exhibit a range of shapes and sizes, from long and narrow in soaring birds to short and rounded in those requiring rapid takeoff. The presence of a tail, often fan-shaped and used for steering during flight, further distinguishes them. Birds also have distinctive legs and feet, with variations such as webbed feet for swimming or strong talons for hunting. Their eyes are generally large and positioned on the sides of their heads, offering a wide field of vision. These combined features create a visual profile unique to birds, setting them apart from other animal categories. This paragraph outlines the visual features of bird that distinguish it from other similar categories. Please use distinguishing visual features to segment bird in the query image.

Input Episode
Cow


Support Image


Query Image


DSV-LFS

Complete class description for mutli-modal LLM for cow object class:

<Qimage>. this one is a query image. Cows possess several distinguishing visual features that set them apart from similar object categories. They have a large, robust body with a pronounced rectangular shape, supported by four sturdy legs ending in cloven hooves. Their heads are relatively large, with broad, flat foreheads and distinctive long, broad snouts. A cow's eyes are large and round, usually positioned on the sides of their head, giving them a wide field of vision. They have large, prominent ears that can be either upright or slightly drooping. One of the most notable features is their pair of horns, which can vary in size and shape but are typically curved and symmetrical, though some cows may be naturally polled (hornless). Their tails are long and thin, ending in a tuft of hair, used to swat away insects. The skin of cows is covered in short hair, with color patterns that can vary significantly, including solid colors, spots, and patches in hues of black, white, brown, or a combination thereof. Unlike other similar animals, cows have a prominent udder with visible teats, particularly in dairy breeds, which is a key distinguishing feature. Additionally, cows have a distinctive gait and posture, often appearing more slow-moving and deliberate compared to other livestock. This paragraph outlines the visual features of cow that distinguish it from other similar categories. Please use distinguishing visual features to segment cow in the query image.

Input Episode
Hair dryer


Support Image


Query Image


DSV-LFS

Complete class description for mutli-modal LLM for hair dryer object class:

<Qimage>. this one is a query image. A hair dryer can be visually distinguished from similar objects primarily by its specific design features. Typically, a hair dryer consists of a cylindrical or slightly tapered body with a prominent handle and a nozzle at one end. The body often features a perforated grill or vents for airflow, which is essential for its function. The handle is ergonomically designed for grip and control, often contrasting in texture or color from the main body to enhance usability and visibility. On the body, there are frequently control buttons or switches for adjusting heat and airflow settings, which are clearly marked and distinct in appearance. The nozzle itself is narrow and elongated, sometimes with a distinct shape or curvature depending on the model, facilitating directional airflow during use. These visual characteristics collectively differentiate a hair dryer from other similar objects like handheld vacuum cleaners or electric razors, which have different body shapes, nozzle configurations, and control mechanisms tailored to their respective functions. This paragraph outlines the visual features of hair dryer that distinguish it from other similar categories. Please use distinguishing visual features to segment hair dryer in the query image.

Chapter 5

Conclusion

In conclusion, this thesis has significantly advanced the state-of-the-art in few-shot semantic segmentation (FSS) by addressing critical challenges associated with low-data regimes. The inherent limitations of conventional FSS methods—such as overfitting to limited support examples, poor generalization to appearance shifts, occlusions, and viewpoint discrepancies—were comprehensively tackled through two innovative frameworks.

Initially, we introduced a transductive meta-learning approach, strategically leveraging an ensemble of features from pretrained classification and semantic segmentation networks. This novel ensemble enhanced the discriminative power by effectively capturing both high-level semantic information and pixel-level spatial details. The proposed two-pass correlation mechanism significantly improved intra-class and intra-object similarity modeling, demonstrating remarkable efficiency by maintaining minimal trainable parameters. Extensive experiments validated its superior performance on standard benchmarks, achieving state-of-the-art results on Pascal-5i and COCO-20i datasets.

Building upon these insights, we subsequently developed DSV-LFS, a pioneering vision-language system that unifies dense visual features with rich semantic knowledge from large multimodal language models (LLMs). By dynamically generating class-specific semantic prompts and integrating them with dense visual correspondences, this framework robustly addresses challenges such as severe appearance variations and ambiguous contexts. The comprehensive evaluations highlighted its outstanding ability to generalize across domains, setting new benchmarks in cross-domain few-shot

segmentation tasks.

Together, these contributions not only elevate the efficacy of few-shot segmentation but also provide a robust foundation for future research in adaptive and intelligent vision systems. Future avenues could explore deeper integration of multimodal knowledge sources, further refinement of transductive learning paradigms, and extending these methodologies to broader applications in computer vision, including medical imaging, and autonomous navigation

References

B., Z., J., X., & T., Q. (2021). Self-guided and cross-guided learning for few-shot segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition.*

Badrinarayanan, V., Handa, A., & Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint arXiv:1505.07293*.

Bilal, A., Sun, G., Mazhar, S., Imran, A., & Latif, J. (2022). A transfer learning and u-net-based automatic detection of diabetic retinopathy from fundus images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 663-674. doi: 10.1080/21681163.2021.2021111

Bilal, A., Zhu, L., Deng, A., Lu, H., & Wu, N. (2022). Ai-based automatic detection and classification of diabetic retinopathy using u-net and deep learning. *Symmetry*.

Boudiaf, M., Kervadec, H., Imtiaz Masud, Z., & Piantanida, P. (2021). Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proceedings of the ieee conference on computer vision and pattern recognition.*

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. In *Advances in neural information processing systems (neurips)* (Vol. 33, pp. 1877–1901).

Brown, T., et al. (2023). Multimodal few-shot learning with frozen language models. In *Advances in neural information processing systems (neurips).*

Bucher, M., Tuan-Hung, V., Cord, M., & Pérez, P. (2019). Zero-shot semantic segmentation. In *Advances in neural information processing systems* (pp. 468–479).

Chen, H., Dong, Y., Lu, Z., Yu, Y., & Han, J. (2022). Pixel matching network for cross-domain few-shot segmentation. In *Proceedings of the european conference on computer vision (eccv).*

Chen, W.-Y., Liu, Z. K., Frank Wang, Y., & Huang, J.-B. (2019). A closer look at few-shot classification. *ICLR*.

Chen, X., Sun, Y., Yan, R., et al. (2023). Unified language model for few-shot text classification and generation. In *International conference on learning representations (iclr).*

Dhillon, G. S., Chaudhari, P., Ravichandran, A., & Soatto, S. (2019). . a baseline for fewshot image classification. *ICLR*.

Dong, N., & Xing, E. P. (2018). Few-shot semantic segmentation with prototype learning. In *Bmvc* (Vol. 3).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fan, Q., Pei, W., Tai, Y.-W., & Tang, C.-K. (2022). Self-support few-shot semantic segmentation. In *European conference on computer vision* (pp. 701–719).

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*.

Gidaris, S., & Komodakis, N. (2018). Dynamic few-shot visual learning without forgetting. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4367–4375).

Guo, Y., Codella, N., Karlinsky, L., V. Codella, J., R. Smith, J., Saenko, K., ... Feris, R. (2020). A broader study of cross-domain few-shot learning. *In European Conference on Computer Vision (ECCV)*.

H., C., D., W., K., M., S., G., & Y., Z. (2021). A unified framework for generalized low-shot medical image segmentation with scarce data. *IEEE Transactions on Medical Imaging*, *40*(10), 2656–2671.

Hariharan, B., & Girshick, R. (2017). Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the ieee intl. conf. on computer vision* (pp. 3018–3027).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In

Proceedings of the ieee conference on computer vision and pattern recognition (cvpr).

He, S., Ding, H., & Jiang, W. (2023). Primitive generation and semantic-related alignment for universal zero-shot segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11238–11247).

Hong, S., Cho, S., Nam, J., Lin, S., & Kim, S. (2021). Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European conference on computer vision, eccv.*

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., ... others (2021). Lora: Low-rank adaptation of large language models. In *International conference on learning representations.*

Hu, T., Yang, P., Zhang, C., Yu, G., Mu, Y., & Snoek, C. G. (2019). Attention-based multi-context guiding for few-shot semantic segmentation. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 8441–8448).

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... Girshick, R. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In *36th conference on neural information processing systems (neurips 2022)* (Vol. 35, pp. 22199–22213).

Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., & Jia, J. (2024). Lisa: Reasoning segmentation via large language model. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Lang, C., Cheng, G., Tu, B., & Han, J. (2022). Learning what not to segment: A new perspective on few-shot segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8057–8067).

Li, A., Luo, T., Lu, Z., Xiang, T., & Wang, L. (2019). Large-scale few-shot learning: Knowledge transfer with class hierarchy. In *Proceedings of the ieee conf. on computer vision and pattern recognition* (pp. 7212–7220).

Li, G., Jampani, V., Sevilla-Lara, L., Sun, D., Kim, J., & 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)* (pp. 8334–8343).

Li, Z., Kamnitsas, K., & Glocker, B. (2019). Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation. In *Medical image computing and computer assisted intervention–miccai 2019: 22nd international conference, shenzhen, china, october 13–17, 2019, proceedings, part iii 22* (pp. 402–410).

Liu, F., Liu, Y., Kong, Y., Xu, K., Zhang, L., Yin, B., ... Lau, R. (2023). Referring image segmentation using text supervision. In *Proceedings of the ieee/cvf international conference on computer vision (iccv)* (pp. 22124–22134).

Liu, H., Jiang, Y., Li, C., et al. (2023). Improved few-shot learning with cross-modal prompt tuning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr).*

Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr).*

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. In *Advances in neural information processing systems (neurips).*

Liu, W., Zhang, C., Ding, H., Hung, T.-Y., & Lin, G. (2022). Few-shot segmentation with optimal transport matching and message flow. In *Proceedings of the european conference on computer vision (eccv).*

Liu, W., Zhang, C., Lin, G., & Liu, F. (2020). Crnet: Cross-reference networks for few-shot segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 4165–4173).

Liu, Y., Lee, J., & Park, M. (2019). Learning to propagate labels: Transductive propagation network for few-shot learning. In *International conference on learning representations(iclr 2019).*

Liu, Y., Liu, N., Cao, Q., Yao, X., Han, J., & Shao, L. (2022). Learning non-target knowledge for few-shot semantic segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 11573–11582).

Liu, Y., Liu, N., Yao, X., & Han, J. (2022). Intermediate prototype mining transformer for few-shot semantic segmentation. *Advances in Neural Information Processing Systems*, *35*, 38020–38031.

Liu, Y., Zhang, X., Zhang, S., & He, X. (2020). Part-aware prototype network for few-shot semantic segmentation. *arXiv preprint arXiv:2007.06309*.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)* (pp. 3431–3440).

Ma, C., Yang, Y., Ju, C., Zhang, F., Zhang, Y., & Wang, Y. (2023). Attrseg: Open-vocabulary semantic segmentation via attribute decomposition-aggregation. In *Proceedings of the neural information processing systems (neurips).*

Masud Ziko, I., Dolz, J., Granger, E., & Ben Ayed, I. (2020). . laplacian regularized few-shot learning. *In International Conference on Machine Learning (ICML).*

Min, J., Kang, D., & Cho, M. (2021). Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 6941–6952).

Moon, S., Sohn, S. S., Zhou, H., Yoon, S., Pavlovic, V., Khan, M. H., & Kapadia, M. (2023). Msi: Maximize support-set information for few-shot segmentation. In *Proceedings of the ieee/cvf international conference on computer vision (iccv)* (pp. 19266–19276).

Nie, J., Xing, Y., Zhang, G., Yan, P., Xiao, A., Tan, Y.-P., ... Lu, S. (2024). Cross-domain few-shot segmentation via iterative support-query correspondence mining. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 3380-3390).

Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the ieee international conference on computer vision* (pp. 1520–1528).

Okazawa, A. (2022). Interclass prototype relation for few-shot segmentation. In *European conference on computer vision, eccv.*

Peng, B., Tian, Z., Wu, X., Wang, C., Liu, S., Su, J., & Jia, J. (2023). Hierarchical dense correlation distillation for few-shot segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 23641–23651).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the international conference on machine learning (icml).*

Rakelly, K., Shelhamer, E., Darrell, T., Efros, A., & Levine, S. (2018a). Conditional networks for few-shot semantic segmentation. *ICLR Workshop track*.

Rakelly, K., Shelhamer, E., Darrell, T., Efros, A. A., & Levine, S. (2018b). Few-shot segmentation

propagation with guided networks. *arXiv preprint arXiv:1806.07373*.

Ravi, S., & Larochelle, H. (2016). Optimization as a model for few-shot learning. *ICLR 2017*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640v5*.

Redmon, J., & Farhadi, A. (2016). Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*.

Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Nips* (p. 91-99).

Rodrıguez, P., Laradji, I., Drouin, A., & Lacoste, A. (2020). Embedding propagation: Smoother manifold for few-shot classification. In *European conference on computer vision, eccv.*

Rusu, A. A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., & Hadsell, R. (2018). Meta-learning with latent embedding optimization. In *Intl. conf. on learning representations.*

Satorras, V. G., & Estrach, J. B. (2018). Few-shot learning with graph neural networks. In *Intl. conf. on learning representations.*

Schwartz, E., Karlinsky, L., Feris, R., Giryes, R., & Bronstein, A. M. (2019). Baby steps towards few-shot learning with multiple semantics. In *Proceedings of the ieee conf. on computer vision and pattern recognition workshop.*

S. Dhillon, G., Chaudhari, P., & Ravichandran, A. (2020). A baseline for few-shot image classification. In *International conference on learning representations(iclr 2020).*

Shaban, A., Bansal, S., Liu, Z., Essa, I., & Boots, B. (2017). One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*.

Shi, X., Wei, D., Zhang, Y., Lu, D., Ning, M., Chen, J., ... Zheng, Y. (2022). Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *Workshop on statistical learning in computer vision, eccv.*

Siam, M., Oreshkin, B. N., & Jagersand, M. (2019). Amp: Adaptive masked proxies for few-shot segmentation. In *Proceedings of the ieee international conference on computer vision* (pp. 5249–5258).

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in neural information processing systems* (pp. 4077–4087).

Sun, G., Liu, Y., Liang, J., & Van Gool, L. (2021). Boosting few-shot semantic segmentation with transformers. *arXiv preprint arXiv:2108.02266*.

Sun, Y., Chen, Q., He, X., Wang, J., Feng, H., Han, J., ... Wang, J. (2022). Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning. *arXiv preprint arXiv:2206.06122*.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., & Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1199–1208).

Team, O. R. (2023). Large language models are good few-shot learners for low-shot image classification. In *Proceedings of the cvpr.*

Teed, Z., & Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision* (pp. 402–419).

Tian, Y., Wang, Y., Krishnan, D., & Tenenbaum, P., Joshua B andIsola. (2020). Rethinking few-shot image classification: a good embedding is all you need? *In European Conference on Computer Vision (ECCV)*.

Tian, Z., Cui, J., Jiang, L., Qi, X., Lai, X., Chen, Y., ... Jia, J. (2023). Learning context-aware classifier for semantic segmentation. In *Proceedings of the thirty-seventh aaai conference on artificial intelligence.*

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... others (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. In *Advances in neural information processing systems* (pp. 3630–3638).

Wang, J., Liu, Y., Zhou, Q., & Wang, F. (2024). Language-guided few-shot semantic segmentation. In *Icassp 2024-2024 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5035–5039).

Wang, K., Liew, J. H., Zou, Y., Zhou, D., & Feng, J. (2019). Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the ieee intl. conf. on computer vision* (pp. 9197–9206).

Wang, W., Duan, L., Wang, Y., En, Q., Fan, J., & Zhang, Z. (2022). Remember the difference:

Cross-domain few-shot semantic segmentation via meta-memory transfer. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 7065–7074).

Wang, Y., Sun, R., & 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)* (pp. 7183–7192).

Wang, Y., Sun, R., Zhang, Z., & Zhang, T. (2022). Adaptive agent transformer for few-shot segmentation. In *Computer vision–eccv 2022: 17th european conference* (pp. 36–52). Springer.

Wang, Y.-X., Girshick, R., Hebert, M., & Hariharan, B. (2018). Low-shot learning from imaginary data. In *Proceedings of the ieee conf. on computer vision and pattern recognition* (pp. 7278–7286).

Wei, J., et al. (2022). Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*. Retrieved from https://arxiv.org/abs/2205.11916

Wu, Z., Shi, X., Lin, G., & Cai, J. (2021). Learning meta-class memory for few-shot semantic segmentation. In *Proceedings of the ieee/cvf international conference on computer vision (iccv)* (pp. 517–526).

Xian, Y., Choudhury, S., He, Y., Schiele, B., & Akata, Z. (2019). Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 8256–8265).

Xie, G.-S., Liu, J., Xiong, H., & Shao, L. (2021). Scale-aware graph neural network for few-shot semantic segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 5475–5484).

Xu, Q., Zhao, W., Lin, G., & Long, C. (2023). Self-calibrated cross attention network for few-shot segmentation. In *Proceedings of the ieee/cvf international conference on computer vision (iccv)*.

Yang, B., Liu, C., Li, B., Jiao, J., & Ye, Q. (2020). Prototype mixture models for few-shot semantic segmentation. *arXiv preprint arXiv:2008.03898*.

Yang, L., Zhuo, W., Qi, L., Shi, Y., & Gao, Y. (2021). Mining latent classes for few-shot segmentation. In *Proceedings of the ieee/cvf international conference on computer vision (iccv)* (p. 8721-8730).

Yang, Y., Chen, Q., Feng, Y., & Huang, T. (2023). Mianet: Aggregating unbiased instance and general information for few-shot semantic segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 7131–7140).

Z., T., H., Z., M., S., Z., Y., R., L., & J., J. (2020). Prior guided feature enrichment network for few-shot segmentation. *IEEE TPAMI*.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part i 13* (pp. 818–833).

Zhang, C., Lin, G., Liu, F., Yao, R., & Shen, C. (2019). Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the ieee conf. on computer vision and pattern recognition* (pp. 5217–5226).

Zhang, G., Navasardyan, S., Chen, L., Zhao, Y., Wei, Y., Shi, H., et al. (2022). Mask matching transformer for few-shot segmentation. *Advances in Neural Information Processing Systems*, *35*, 823–836.

Zhang, J., Sun, Y., Yang, Y., & Chen, W. (2022). Feature-proxy transformer for few-shot segmentation. *ArXiv*, *abs/2210.06908*.

Zhang, X., Wei, Y., Yang, Y., & Huang, T. S. (2020). Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics*, *50*(9), 3855–3865.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... others (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhao, W. X., Zhou, K., et al. (2023). Harnessing large language models for zero-shot and few-shot learning in knowledge-intensive tasks. *arXiv preprint arXiv:2304.12345*. Retrieved from https://arxiv.org/abs/2304.12345

Zhu, L., Chen, T., Ji, D., Ye, J., & Liu, J. (2024). Llafs: When large language models meet few-shot segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.