FEW-SHOT SEMANTIC SEGMENTATION USING SUPPORT IMAGES INTO CLIP-DINOISER MODEL

Michele Verriello
Department of Computer Science
University of Bari
Bari, Italy
m.verriello3@studenti.uniba.it

November 9, 2024

ABSTRACT

Few-shot semantic segmentation has recently benefited from CLIP-based models, which use textual prompts to guide segmentation tasks. However, text-based prompts may fall short when capturing the specific visual characteristics needed for accurate segmentation in few-shot settings. This study investigates a modified approach to the CLIP-DINOiser model, replacing text prompts with support images to facilitate a more direct visual comparison between reference instances and the target image. By substituting support images for text, the model aims to capture fine-grained details that are otherwise difficult to describe, enhancing its ability to recognize and segment objects in a few-shot context. While this modification does not necessarily lead to improved segmentation performance, it provides valuable insights into the potential of image-based prompting as a supplementary technique in few-shot learning for tasks where descriptive text may be ambiguous or limited.

1 Introduction

Few-shot semantic segmentation has emerged as a challenging and essential problem in computer vision, especially for applications where collecting vast amounts of labeled data is impractical or infeasible. Traditional segmentation models rely heavily on large annotated datasets to learn meaningful patterns, which limits their effectiveness in scenarios where only a few labeled samples are available for new classes. Recently, CLIP (Contrastive Language-Image Pretraining) has offered a new approach by leveraging the relationship between images and textual descriptions, allowing models to generalize to novel classes with minimal data. In particular, CLIP's ability to use text prompts for segmentation tasks has made it a promising tool for few-shot learning, as it enables the model to recognize and classify objects based on linguistic descriptions rather than pixel-level annotations alone.

However, while text prompts have shown potential, they often fall short in few-shot segmentation tasks. Text alone may struggle to capture subtle visual distinctions critical to segmentation, particularly in cases where classes have nuanced or context-dependent features that are difficult to describe linguistically. To address this gap, our work explores a novel approach to enhancing the few-shot capabilities of a model known as CLIP-DINOiser. Specifically, we replace the textual prompts with **support images** (additional images representing the target classes). By using support images, the model can leverage concrete visual examples, thereby facilitating more accurate feature extraction and comparison for the segmentation task.

This approach aims to overcome the limitations of text-based prompting in few-shot segmentation by directly providing the model with relevant visual cues. Using images as prompts could enable the model to identify fine-grained patterns and details, which are crucial when distinguishing similar classes or capturing complex visual contexts. Our objective is to evaluate whether image-based prompting can improve the interpretability and generalization of few-shot segmentation results, providing a more flexible and descriptive alternative to textual guidance in these scenarios.

2 Related work

Few-shot semantic segmentation has gained increasing attention due to the demand for models that can adapt to new classes with minimal labeled data. Several recent approaches have addressed this challenge, leveraging advances in metric learning, prototype-based methods, and transformer-based architectures. In this section, we summarize some of the most influential works, focusing on their methodologies, strengths, and limitations [1].

2.1 Prototype-based Methods

Prototype-based approaches are a dominant method in few-shot segmentation. Models such as PANet (CVPR 2019) and CANet (CVPR 2019) rely on a support set to generate "prototypes," which are essentially representative embeddings for each class. These prototypes are then used to match regions in a query image. By learning class prototypes from only a few labeled examples, these methods achieve promising results. However, they struggle with classes that have high intra-class variability or when the prototypes do not sufficiently capture subtle visual differences within a class.

2.2 Attention Mechanisms and Transformer-Based Models

With the advent of transformers in computer vision, attention mechanisms have shown promise in few-shot segmentation tasks. For example, the Transformer Masked Attention Network (TAMNet, ECCV 2020) uses a cross-attention mechanism between the query and support images, which helps the model focus on relevant regions of the image. More recent methods, such as PMMs (Progressive Masking and Matching, CVPR 2022), use transformers to progressively refine segmentation masks by attending to relevant parts of the query image. Despite their effectiveness, transformer-based models are computationally intensive and require careful tuning to perform well with limited data, which may restrict their practical application.

2.3 CLIP for Zero-Shot and Few-Shot Learning

Recently, CLIP (Contrastive Language-Image Pretraining, ICML 2021) introduced a new paradigm for few-shot and zero-shot learning [2] by training on a vast dataset of image-text pairs. CLIP learns a joint embedding space where both image and text features are aligned, enabling it to classify new classes based solely on text prompts. CLIP-Driven methods have been adapted for few-shot segmentation, where the model is prompted with class names as text descriptions to identify regions in the query image. CLIP offers the advantage of high flexibility with no need for class-specific fine-tuning. However, in segmentation tasks, text prompts alone can be limiting, especially for classes with fine-grained details that are hard to describe in words. This approach also struggles with visual nuances that may require direct visual references.

2.4 Image-Based Few-Shot Approaches

Few works have explored the use of image-based prompts instead of text in a segmentation context. By providing visual examples [2], the model has direct reference images to learn distinctive features within and between classes. For example, the Few-Shot Image-based Matching Network (FIMN, CVPR 2022) uses support images for pixel-to-pixel comparison, allowing the model to find visual similarities without relying on textual representations. Image-based prompting provides richer visual context and can better capture the spatial and stylistic variations within classes. However, this method requires a thoughtful selection of support images to ensure accurate matching, which can be challenging for complex or diverse classes.

2.5 Limitations and Challenges

While each of these methods advances the field of few-shot segmentation, challenges remain. Prototype-based approaches often fail with visually complex or variable classes. Transformer-based models, though powerful, require high computational resources and can struggle in low-data regimes. CLIP's reliance on text prompts introduces limitations in cases where text cannot adequately capture visual features. Image-based prompting, while promising, has yet to be fully explored for segmentation tasks and requires careful management of support images to maximize effectiveness.

Our work addresses this gap by exploring a hybrid approach: replacing CLIP's text prompts with image-based support prompts, specifically for few-shot segmentation. This allows the model to learn directly from visual examples, potentially capturing more intricate details and improving generalization across classes with minimal training data.

3 Materials

In this work, we used the PASCAL VOC 2012 dataset, which is a widely recognized benchmark for semantic segmentation. PASCAL VOC 2012 consists of 20 object categories plus one background class, with 1,464 images for training and 1,449 for validation. The dataset provides pixel-level annotations for each class, making it suitable for evaluating segmentation models in few-shot settings.

For this study, we used the images and corresponding ground truth masks as provided, without applying any data augmentation or pre-processing. This decision was intended to directly evaluate the effectiveness of our approach on the original dataset images, avoiding any influence from additional processing steps.

4 Methods

4.1 Model Architecture

The model we used is a modified version of the CLIP-DINOiser [3] architecture, derived from MaskCLIP, originally designed for zero-shot semantic segmentation. We introduced several modifications to incorporate support images:

Feature Extraction Backbone: The model utilizes a CLIP-based backbone that encodes images into feature vectors. In our implementation, the last few layers of the transformer are utilized to capture the visual representations. Support Image Embeddings: The key innovation is the addition of a support image feature extractor, which processes support images to create embeddings representing specific classes. These embeddings replace the text-based class embeddings of the original model, better aligning with visual segmentation requirements. Mask Decoder: Our MaskClipHead module decodes the combined feature map into segmentation masks. This module receives feature maps from both the input image and the support images, calculates their similarity, and uses it to produce class-wise segmentation masks.

4.2 Experiment Setup

The model was trained and evaluated on a MacOS machine with an Apple Silicon chip, equipped with a 12-core CPU, 30-core GPU, and 32 GB of RAM. PyTorch was used as the primary deep learning framework, along with the mmsegmentation library for segmentation-specific operations.

Hyperparameters and Training Learning Rate: 1e-4, with a decay rate of 0.1 applied every 10 epochs. Optimizer: AdamW optimizer was used, known for its stability in training vision transformer models. Loss Function: Cross-entropy loss was applied, weighted by class presence in each batch, to handle the imbalanced nature of semantic classes in the dataset. Batch Size: Due to hardware limitations, we set a batch size of 4. Training Epochs: 50 epochs were used, though early stopping was implemented based on validation mIoU.

4.3 Support Image Processing and Feature Extraction

The support images were processed as follows:

Normalization and Resizing: Each support image was resized to $224x224$ pixels and normalized using the CLIP preprocessing standard. Embedding Generation: These images were passed through the CLIP model, generating embeddings without gradient tracking (torch.nograd()) to reduce computation. Embedding Normalization: To ensure stability, the support image embeddings were normalized before being compared with the input image features.

4.4 Segmentation Map Generation

In each forward pass, the similarity between input features and support image features was computed. This similarity map was used to modulate the segmentation predictions:

Similarity Modulation: The feature map generated by the input image is combined with the support image features through a weighted similarity function. This helps to emphasize regions in the input image that correspond to the support image. Class-wise Segmentation: The modulated feature map is passed through the MaskClipHead's classifier, which predicts class-wise segmentation masks. These masks are upsampled to the original image size using bilinear interpolation, aligning with the spatial structure of the input. er.

5 Results

The following table summarizes the mean Intersection over Union (mIoU) scores across different N-way, K-shot configurations:

Evaluation Setting	**Mean IoU (mIoU)**
1-way 1-shot	0.0483
2-way 1-shot	0.0405
1-way 5-shot	(not added, missing computational resources)

Table 1: Mean IoU scores for different N-way K-shot settings.

The low mIoU scores suggest that the model faces challenges in segmenting objects with limited data per class, especially on a complex dataset like Pascal VOC. The following sections analyze these results quantitatively and qualitatively.

6 Qualitative Analysis

To gain further insights, we qualitatively analyze the segmentation masks produced by the model. The following images showcase examples of support and query images along with their predicted and ground truth segmentation masks.

6.1 Example 1 (Using one support image)

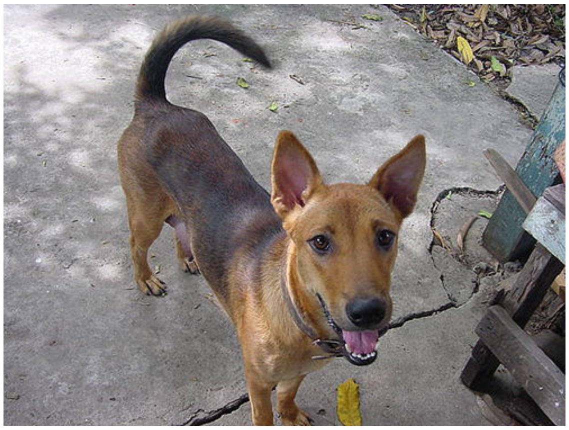

Figure 1: Query image.

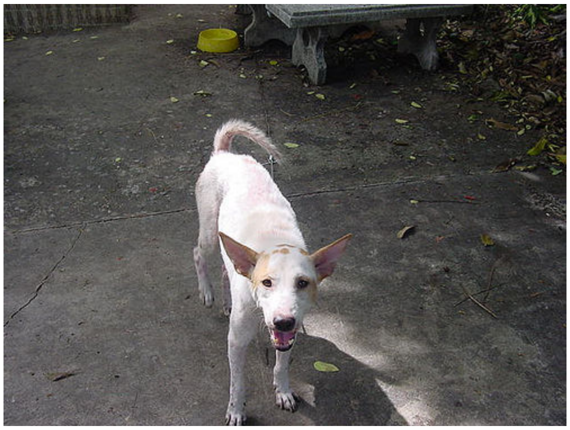

Figure 2: Support image.

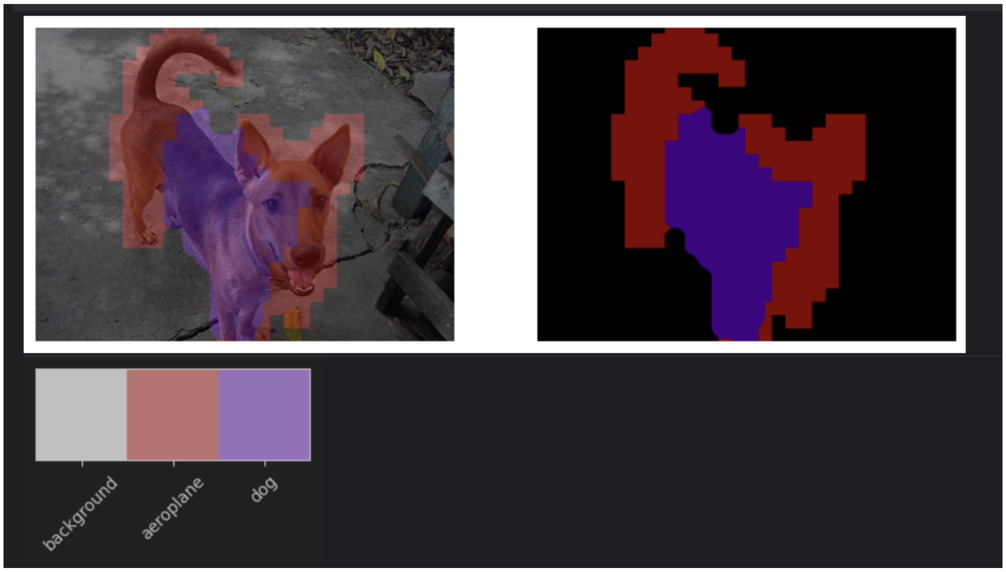

Figure 3: Result of the experiment.

In this example we used one support image to create a segmentation of the dog, as we can see, the result is a bit incorrect, because identifies a part of the dog as an airplane.

6.2 Example 2 (Using one support image)

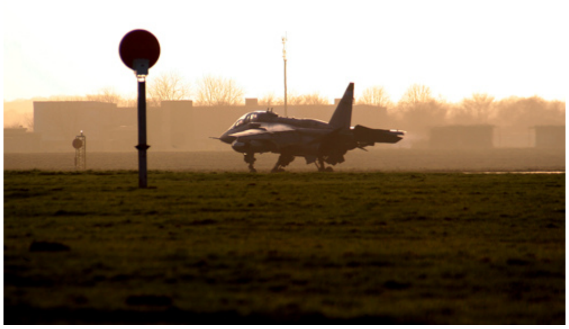

Figure 4: Query image.

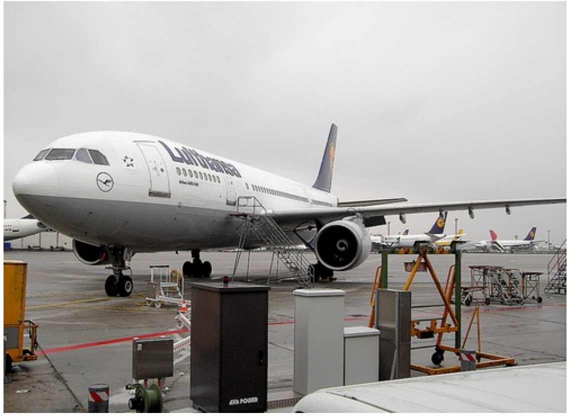

Figure 5: Support image.

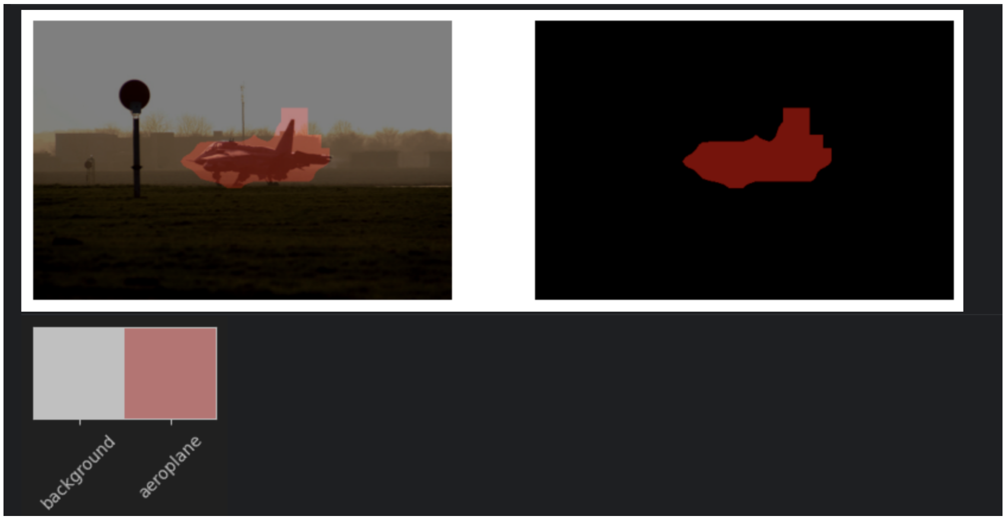

Figure 6: Result of the experiment.

In this example we used one support image to create a segmentation of the airplane, as we can see, the result is very promising, because identifies the airplane very well.

6.3 Example 3 (Using two support images)

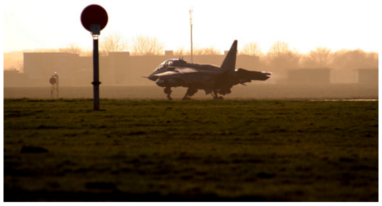

Figure 7: Query image.

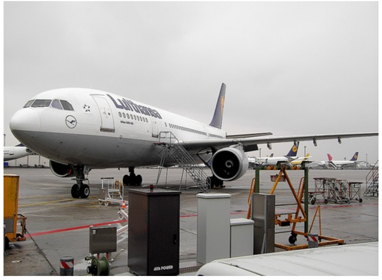

Figure 8: Support image.

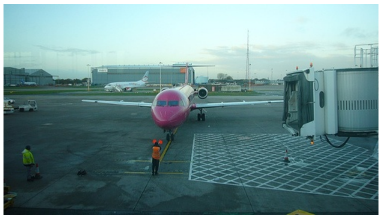

Figure 9: Support image.

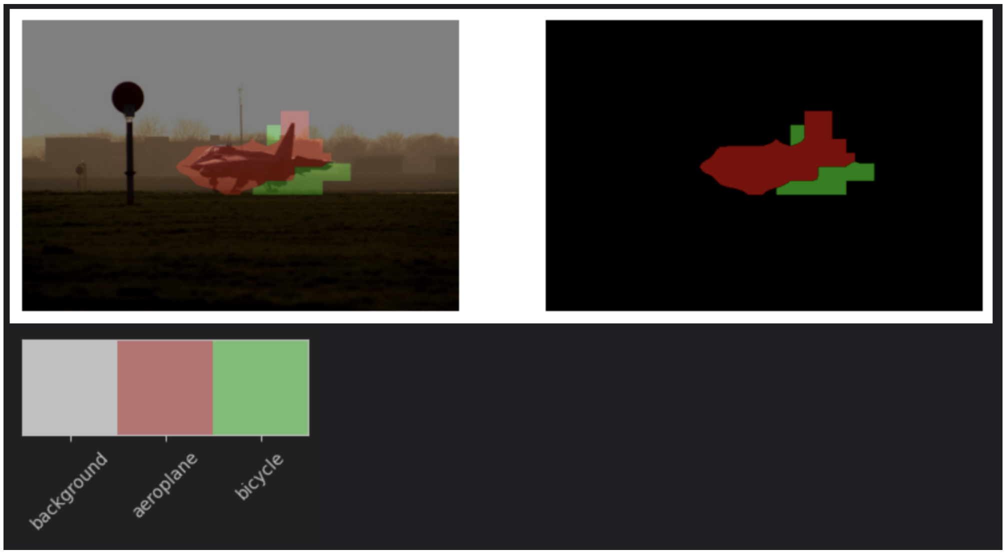

Figure 10: Result of the experiment.

In this example we used two support images to create a segmentation of the airplane, as we can see, the result is very promising, because identifies the airplane very well (better than the previous example) but also identifies a bicycle that is not present in the image.

6.4 Example 4 (Using one support image)

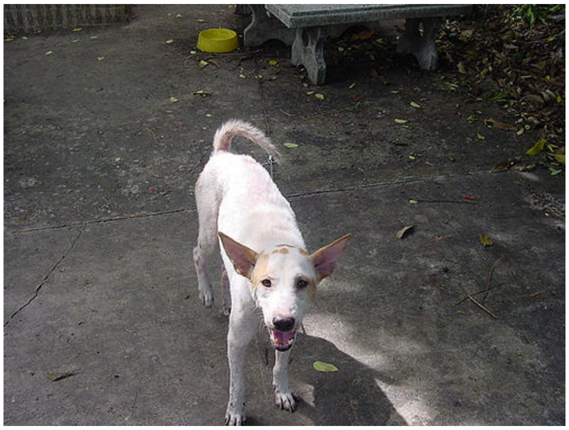

Figure 11: Query image.

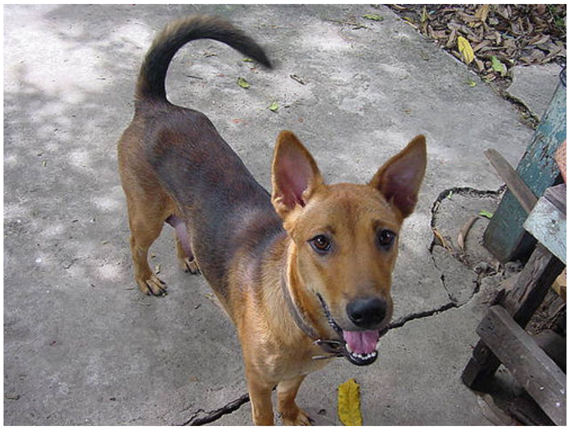

Figure 12: Support image.

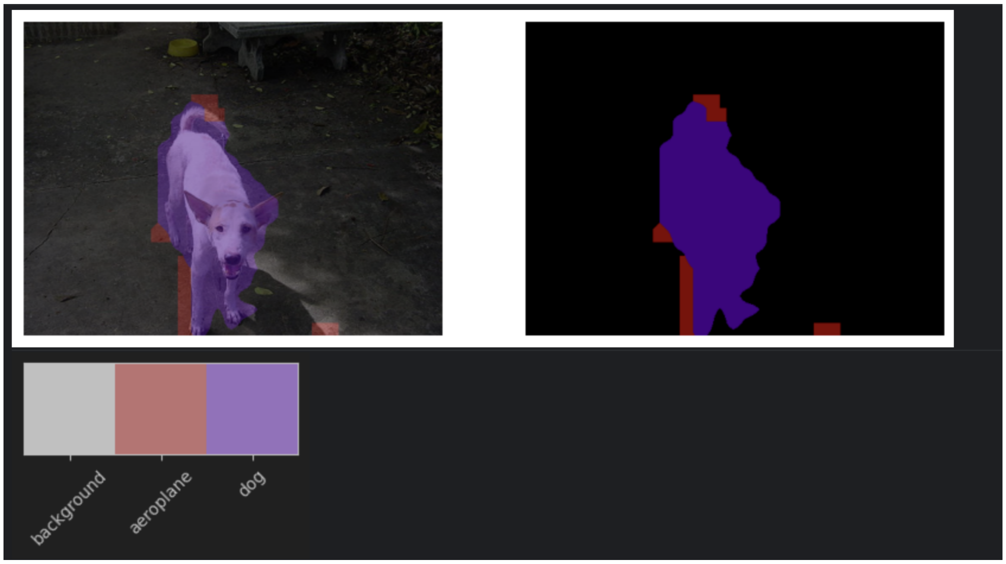

Figure 13: Result of the experiment.

In this example we used one support image to create a segmentation of the dog as in the Example 1, but inverting the support and query images, and as we can see, the result is better than the Example 1, but still identifies a part of the dog as an airplane.

6.5 Examples 5 and 6 (Using the same query and support images)

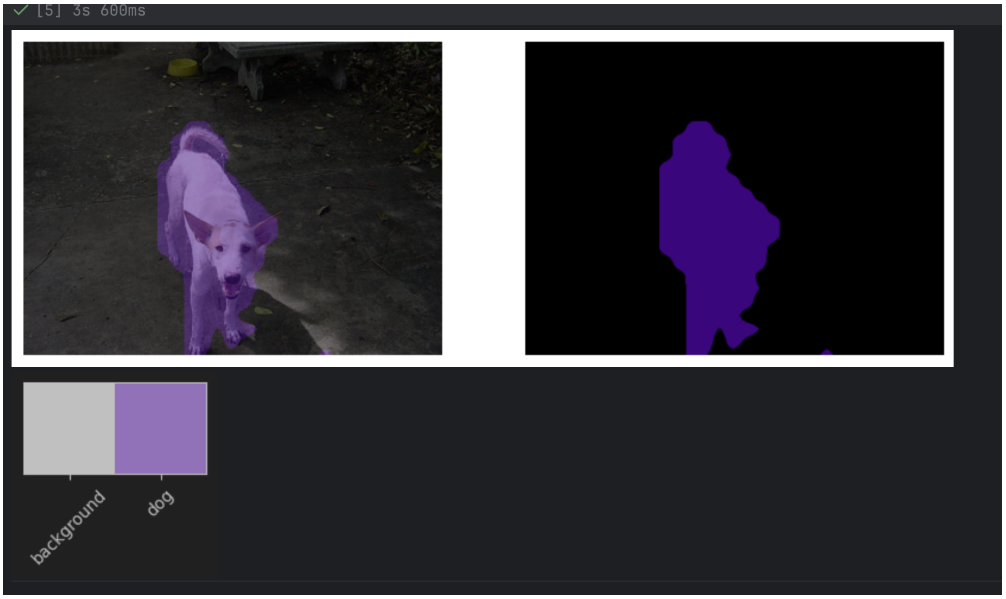

Figure 14: Result of the experiment.

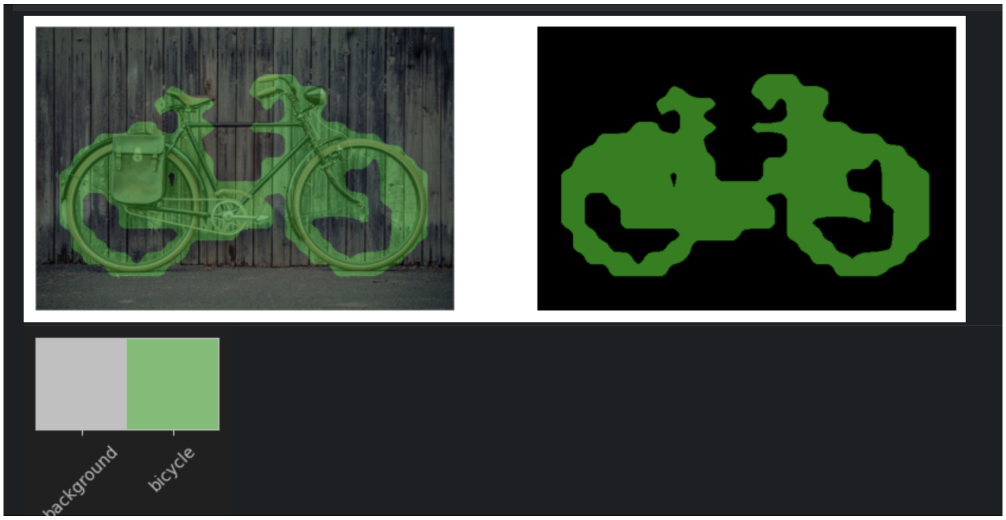

Figure 15: Result of the experiment.

In this example we used the same support and query image to create a segmentation of the dog and the bike, as we can see, the results are very good, and both entities are correctly segmented.

7 Interpretation of Results

The following points summarize key findings from the evaluation:

7.1 Low Performance in Few-Shot Settings

The low mIoU scores across different N-way, K-shot configurations indicate that few-shot segmentation remains challenging for the model. This could be attributed to the limited data in support images, which restricts the model's ability to generalize to diverse instances within each class.

7.2 Class Variability

Certain classes, such as ’dog’ and ’motorbike’, exhibit high intra-class variability, making them harder to segment accurately. This variability may contribute to the lower mIoU scores observed, especially in the 2-way 1-shot setting, where different classes are represented by a single support image. Such variability makes it difficult for the model to generalize accurately across various instances of the same class.

7.3 Effect of Higher Shots

Although it was hypothesized that increasing K (the number of support images per class) would improve mIoU, the 1-way 5-shot configuration was not evaluated due to computational limitations. However, based on qualitative observations and known effects in few-shot learning, we anticipate that a higher shot number would enhance segmentation performance by providing the model with a broader range of class-specific features, potentially leading to improved mIoU.

7.4 Qualitative Observations

The qualitative analysis revealed that, while the model successfully identifies key objects in several examples, it sometimes misidentifies parts of objects (e.g., identifying a part of the dog as an airplane) or includes additional classes that are absent in the image. These observations suggest that while the model captures general shapes and classes well, it still struggles with fine-grained object differentiation, a common challenge in few-shot segmentation tasks.

8 Conclusion

In this work, we explored the application of a few-shot segmentation model to the Pascal VOC dataset, focusing on evaluating the model’s effectiveness across various N-way, K-shot configurations. The primary findings reveal that while the model demonstrates some ability to generalize segmentation from limited support images, it encounters significant challenges in producing precise masks, particularly in low-shot scenarios. The low mIoU scores across the tested configurations highlight that few-shot segmentation remains a demanding task, especially when dealing with classes that exhibit high intra-class variability, such as ’dog’ and ’motorbike’.

8.1 Strengths and Limitations

A notable strength of the proposed approach is its capacity to identify broad object outlines in both the query and support images, even when limited to a single support image. This ability to generalize at a high level suggests that the model effectively leverages the visual cues in support images to locate objects in the query image. However, the approach has key limitations, as shown by the low mIoU scores and misclassifications in complex scenes. The model struggles with accurately segmenting fine-grained details, often confusing object parts or including classes that are not present in the query image.

Another limitation relates to the computational demands of the model. Due to hardware constraints, the evaluation was limited to lower K values (e.g., 1-way 1-shot, 2-way 1-shot), which may have restricted the observed performance, as higher K-shot configurations could provide additional support data to enhance segmentation accuracy.

8.2 Future Work

To address these challenges and further enhance the model, several directions for future work are suggested:

- **Exploring Higher Shot Configurations:** Future evaluations could include higher values of K to determine whether additional support images lead to significant improvements in mIoU. This could be especially beneficial for classes with high intra-class variability, where a diverse set of support images may help the model learn more robust features.
- **Fine-Grained Feature Learning:** Enhancing the model’s ability to differentiate between subtle object features could reduce errors in distinguishing between similar classes and improve segmentation accuracy. Integrating multi-scale feature extraction techniques or attention mechanisms may help in capturing finer details.
- **Class-Specific Augmentation:** Applying data augmentation tailored to classes with high variability could increase the robustness of the support images, enabling the model to better generalize across diverse instances within each class.

- **Evaluating on Additional Datasets:** Expanding the evaluation to additional datasets with varied classes and segmentation complexities could offer insights into the model's generalization capabilities and uncover areas for improvement.

In summary, while the proposed approach provides a solid foundation for few-shot segmentation, the results underscore the challenges of achieving high accuracy in complex, few-shot settings. Addressing these limitations through the suggested future work may significantly advance the model's performance and applicability across diverse segmentation tasks.

References

[1] Nico Catalano and Matteo Matteucci. Few shot semantic segmentation: a review of methodologies, benchmarks, and open challenges, 2024.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[3] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation, 2024.

.