

Few Shot Semantic Segmentation: a review of methodologies, benchmarks, and open challenges

NICO CATALANO and MATTEO MATTEUCCI, Politecnico di Milano, Italy

Semantic segmentation, vital for applications ranging from autonomous driving to robotics, faces significant challenges in domains where collecting large annotated datasets is difficult or prohibitively expensive. In such contexts, such as medicine and agriculture, the scarcity of training images hampers progress. Introducing Few-Shot Semantic Segmentation, a novel task in computer vision, which aims at designing models capable of segmenting new semantic classes with only a few examples. This paper consists of a comprehensive survey of Few-Shot Semantic Segmentation, tracing its evolution and exploring various model designs, from the more popular conditional and prototypical networks to the more niche latent space optimization methods, presenting also the new opportunities offered by recent foundational models. Through a chronological narrative, we dissect influential trends and methodologies, providing insights into their strengths and limitations. A temporal timeline offers a visual roadmap, marking key milestones in the field's progression. Complemented by quantitative analyses on benchmark datasets and qualitative showcases of seminal works, this survey equips readers with a deep understanding of the topic. By elucidating current challenges, state-of-the-art models, and prospects, we aid researchers and practitioners in navigating the intricacies of Few-Shot Semantic Segmentation and provide ground for future development.

CCS Concepts: • Computing methodologies → Image segmentation; Machine learning.

Additional Key Words and Phrases: few shot learning, prototypical networks, conditional networks, latent space optimization, contrastive learning, generative models, variational auto encoders

ACM Reference Format:

Nico Catalano and Matteo Matteucci. 2018. Few Shot Semantic Segmentation: a review of methodologies, benchmarks, and open challenges. *J. ACM* 37, 4, Article 111 (August 2018), 32 pages. <https://doi.org/XXXXXX.XXXXXXXX>

1 INTRODUCTION

Computer Vision (**CV**) plays a crucial role in various application fields, from robotics and clinical analysis to video surveillance; semantic segmentation represents one of the most significant tasks in this research field. Semantic segmentation can be described as predicting pixel-level category labels, thereby generating a segmentation mask for fixed subject categories in a given image [32, 34, 61]. To better contextualize the role of semantic segmentation, it is essential to understand its relationships with other **CV** tasks such as image classification, object detection, parts segmentation, etc. Image classification focuses on comprehending the overall scene by assigning one or more labels to an entire image (see Figure 1a). Object detection (see Figure 1b) concentrates on predicting object locations, typically supplying bounding boxes for the identified objects. Parts segmentation (depicted in Figure 1d) closely resembles semantic segmentation, aiming to predict pixel-level segmentation masks that cover distinct parts constituting the

Authors' address: Nico Catalano, nico.catalano@polimi.it; Matteo Matteucci, matteo.matteucci@polimi.it, Politecnico di Milano, Piazza Leonardo DaVinci 32, Milano, Italy, 20133.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Association for Computing Machinery.

Manuscript submitted to ACM

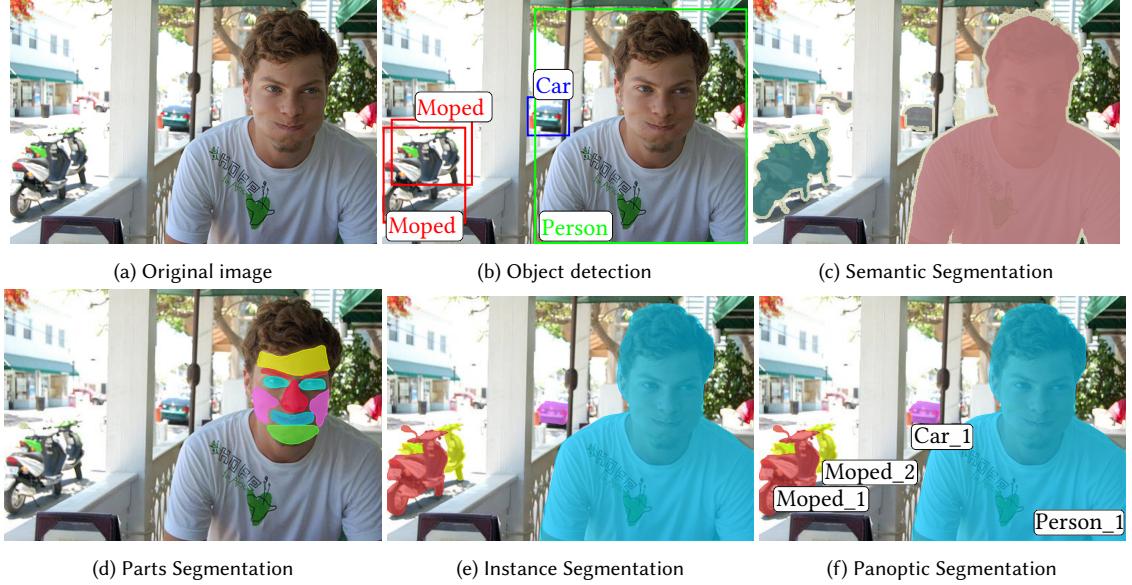


Fig. 1. Comparison of various Computer Vision (CV) tasks. In a given image (depicted in picture a), there are multiple layers of interpretation possible. Object Detection (picture b) identifies the region containing a subject using bounding boxes. Semantic Segmentation (picture c) assigns a class label to each pixel without necessarily distinguishing different instances of the same subject class. Parts Segmentation (picture d) provides pixel-level annotation by segmenting the constituent parts of the intended subject, such as the face parts in this example. Instance Segmentation (picture e) distinguishes the subjects in the scene with a different segmentation mask for each one but does not necessarily assign a semantic class. Panoptic Segmentation (picture f) involves the semantic classification of every pixel in an image, coupled with the delineation of each unique subject instance mask. In the given image, two mopeds are identified with labels “Moped_1” and “Moped_2”, illustrating the distinct instances within the same class.

intended subject. Instance segmentation (see Figure 1e) aims at differentiating individual subjects in an image, even if they share the same kind, without necessarily assigning them a category. Finally, panoptic segmentation (Figure 1f) seamlessly combines semantic segmentation with instance segmentation, predicting pixel-level categories and distinguishing each class instance in the scene. Figure 1, which visually compares various CV tasks, puts semantic segmentation as a crucial intermediary task, as it provides enough image understanding to enable a large number of downstream applications but still lacks deeper analysis and interpretation.

The advent of Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN) of recent years has revolutionized the Artificial Intelligence (AI) community and the Computer Vision (CV) research field allowing the introduction of semantic segmentation models capable of impressive accuracy, with works by Long et al. [76], Badrinarayanan et al. [5], Ronneberger et al. [94], Chen et al. [18], and Zhao et al. [138] underscoring the potential of these architectures. However, a significant drawback emerges for semantic segmentation models: their reliance on extensive training datasets.

Creating and labeling large datasets for semantic segmentation can demand substantial resources. For example, labeling object instances for 80 classes in 328,000 images for the MS COCO [66] dataset required 55,000 worker hours. This intensive data collection process poses significant financial and temporal challenges, especially in domains like agriculture or healthcare, where data variability and scarcity, requirements of domain-specific knowledge, and privacy concerns hinder acquisition efforts. Indeed, the desire to develop models capable of generalizing from a limited number

of training samples is not new to the **Artificial Intelligence (AI)** and **Machine Learning (ML)** communities and falls under the definition of **Few Shot Learning (FSL)** [44, 49, 92, 116]. Further specifying the **FSL** problem to the semantic segmentation task takes the name of **Few Shot Semantic Segmentation (FSS)**, and it is the focus of our contribution.

In particular, this work discusses the main approaches and strategies adopted to design models capable of learning to segment a new semantic class with just few labelled examples (e.g., from 1 to 5). Our examination navigates the main families of key solutions proposed in the **FSS** literature and describes commonly used public benchmark datasets that serve as crucial evaluation tools for **FSS** models. Additionally, we present and compare the outcomes of noteworthy works in the field, providing both quantitative results on benchmark datasets as well as selected samples of predictions, providing this way also qualitative results. Furthermore, building upon the insights gained from previous surveys [15, 92], we extend our discussion to encompass the outlooks and future directions of **FSS**. This includes an examination of the rapidly expanding and promising innovations introduced with vision foundational models. In our conclusion, we address primary critiques and potential extensions, offering a nuanced overview of the current landscape and illuminating potential future trajectories in this evolving field.

A key distinction of our work from previous surveys, such as that of Ren et al. [92], lies in its focused scope. While Ren et al. explored the broader scope few and zero-shot visual semantic segmentation methods across 2D and 3D spaces, this survey is dedicated exclusively to **FSS**, providing a more updated and detailed analysis specifically tailored to the segmentation task in the few-shot scenario. This focused approach allows for a deeper understanding and more comprehensive evaluation of techniques relevant to this particular area. Furthermore, our survey paper fills a critical gap in the literature by offering a specialized examination of **FSS**. Although previous surveys have touched upon related topics, such as [15, 92], the rapid advancements and unique challenges within the realm of few-shot segmentation necessitate dedicated attention. By focusing solely on this niche, we are able to provide nuanced insights, identify emerging trends, and highlight key areas for future research.

2 BACKGROUND AND PROBLEM DEFINITION

Few Shot Semantic Segmentation (FSS) integrates principles from **Few Shot Learning (FSL)** into the domain of semantic segmentation by specializing the semantic segmentation task to a low data regime. To elucidate the shared concepts among **FSS** and **FSL**, we commence by recalling the fundamental definition of **Machine Learning (ML)**, as articulated by Mitchell et al. [82] and Mohri et al. [83]:

Definition 1 (Machine Learning [82, 83]). A computer program is said to learn from experience E with respect to some classes of task T and performance measure P if its performance can improve with E on T measured by P .

This fundamental notion lays the groundwork for the concept of **Few Shot Learning (FSL)**, where the learning paradigm is focused on scenarios where available experience E is limited. Wang et al. [116] define this ideas as **FSL**:

Definition 2 (Few Shot Learning (FSL) [116]). **Few Shot Learning (FSL)** is a type of machine learning problems (specified by E , T , and P), where E contains only a limited number of examples with supervised information for the target T .

This transition from the broader landscape of **ML** to the nuanced realm of **FSL** becomes particularly pertinent when applied to the challenges of semantic segmentation. The objective of this task is to predict a segmentation mask over an input image covering the intended class of subject. Models for this task often rely on large-scale training datasets, presenting a significant challenge in scenarios where data availability is limited. Here, the concept of **Few Shot Semantic**

Segmentation (FSS) emerges as a natural progression, seamlessly integrating principles from **FSL**, as articulated by Shaban et al. [96]:

Definition 3 (Few Shot Semantic Segmentation (FSS) [96]). Few Shot Semantic Segmentation (FSS) is the problem of predicting the semantic segmentation mask \hat{M}_q of a subject class l in a query image I_q given a support set S composed by a limited training set of k image mask pairs.

This definition can be further formalized operationally as it follows. Given a semantic class l , we construct a support set of k image-mask pairs:

$$S(l) = \{(I^i, M_l^i)\}_{i=1}^k,$$

where, I^i represents an RGB image of shape $[H^i, W^i, 3]$, and M_l^i is a binary mask of shape $[H^i, W^i]$ segmenting the class l in I^i . Additionally, a query image I_q is considered, and the objective is to learn a model f_θ such that:

$$\hat{M}_q = f_\theta(I_q, S(l)).$$

This model predicts the binary mask \hat{M}_q for the semantic class l in the query image I_q . In this context, the number of available shots k is crucial, leading to the designation of FSS as K-FSS in various works, with k often set to 1 or 5 [47, 74, 89, 96, 105, 119, 132, 136].

Given the scarcity of labelled samples, traditional algorithms can be inadequate for training FSS models. In order to extract maximum learning from the limited number of examples, most studies [7, 24, 62, 78, 84, 89, 96, 98, 105, 113, 119, 132] have employed episodic training which is a form of meta-learning, or learning-to-learn approach. Another common strategy is transfer learning, which capitalizes on the knowledge acquired by pre-training a portion of the model on different tasks or datasets. Being it a basic component, the following section presents an overview of the use of episodic training in FSS and it explains how pre-training can be applied to FSS.

2.1 Episodic training

The main characteristic of FSS is its limited number of labelled examples. This aspect is particularly challenging for training procedures, rendering traditional methods inadequate. To better adapt to the FSS settings, it is common choice to resort to episodic training. Episodic training is a meta-learning methodology where the model is trained through a series of “episodes”. According to this approach, initially proposed by Vinyals et al. in [109], a model can be trained with a series of meta-learning episodes composed by a support set of labeled images S , a query image I_q and its corresponding ground truth mask M_q . The model then learns by minimizing the loss between the predicted mask \hat{M}_q and the ground truth mask M_q for each episode. The model can later be tested with meta-testing episodes on unseen classes.

This concept can be detailed as it follows: given the set C of the classes in the dataset D_{train} , we split it in C_{train} and C_{test} such that $C_{train} \cap C_{test} = \emptyset$. We can now form a training episode by randomly sampling one label class c from C_{train} . Fixed c , we sample k unique images and their corresponding masks segmenting the class c in them to form the support set S , and one image query I_q and its ground truth M_q for the class c . Given the support set S , the query image I_q and the ground truth M_q as described, the training objective of the model is:

$$\hat{\theta} = \arg \max_{\theta} \mathbb{E}_{S, (I_q, M_q)} [\log P_\theta(M_q | I_q, S)] \quad (1)$$

Optimizing θ in Equation 1 implements a form of meta-learning since, in each episode, the model is learning to learn how to predict a segmentation mask for an unseen class given only a limited support of examples. Thus, with enough episodes, the model is expected to learn how to generalize from a limited support set.

2.2 Transfer Learning

In order to maximize the information extraction from limited data, a prevalent strategy within the literature involves the implementation of transfer learning, a set of techniques in which knowledge acquired from one task or dataset is applied to another related domain. In the context of FSS one common implementation of transfer learning is the use of pretrained feature extractors or backbones. These feature extractors can be initially trained to learn general visual representations on large datasets, such as the object detection dataset ImageNet [23]. Leveraging this idea, works such as those by [24, 84, 89, 96, 136] integrate a frozen pretrained feature extractor into their pipelines allowing for the extraction of domain-agnostic visual cues, such as shapes and patterns.

While the reliance on a large dataset for pretraining may seem contradictory to the low-data regime of FSS, it is crucial to recognize that the term ‘few shots’ pertains solely to the new class that the model needs to learn. Therefore, under the condition that the semantic classes utilized in pretraining are distinct from those to be learned with few shots, pretrained feature extractors are a permissible form of transfer learning in this context.

In sum, to fully exploit the few examples given for each new class, a common practice found in the literature is to perform pre-training on an external $D_{pretrain}$ like ImageNet and train on the given dataset D_{train} using episodic learning. In this scheme, a model can have a pre-training dataset $D_{pretrain}$ and a training dataset D_{train} where $D_{pretrain} \neq D_{train}$. In the field of FSS, Episodic learning is implemented by splitting the label set of all classes C in the dataset D_{train} in C_{train} and C_{test} such that $C_{train} \cap C_{test} = \emptyset$. In addition, a portion of the model can be pre-trained on $D_{pretrain}$, and the entire model can undergo meta-train on episodes from D_{train} with label class from C_{train} , and meta-test with episodes from D_{test} on classes from C_{test} .

3 METHODOLOGY SPECTRUM

The field of FSS can be a multifaceted subject that can be approached from various perspectives. This section encompasses some of the most influential and notable studies in this area. To accommodate a diverse range of works and research lines and avoid overly strict categorization, we start identify three main strategies: conditional networks, prototypical learning, and latent space optimization. Latent space optimization can be further specialized as generating synthetic datasets, e.g., by using GAN encodings, and via contrastive learning. Moreover, in this section we present the new opportunities offered by the recent foundation models and how they can be applied to tackle the FSS task. Figure 2 depicts a timeline indicating the most influential FSS models and the definitions of the task. In addition, the proposed timeline reports the trends and influences of the approaches presented in this paper. In the following sections, we describe each approach and highlight the key details of representative works.

3.1 Conditional Networks

The earliest works in the field, by Shaban et al. [96] and Rakelly et al. [89], addressed the FSS challenge by using two branched models as depicted in Figure 3. This kind of models has two parallel branches: the conditioning branch and the segmentation branch. The conditioning branch g takes in input the support set S and produces a parameter set θ . The segmentation branch h gets as input the query image I_q and the parameter set θ generated by the conditioning branch g to produce the prediction mask \hat{M}_q . During testing, the model works by first extracting a dense feature volume from I_q using a pre-trained feature extractor in the segmentation branch h . Then the segmentation branch h utilizes both the feature volume and the parameter set θ to predict the final mask \hat{M}_q . We can model the conditioning branch g

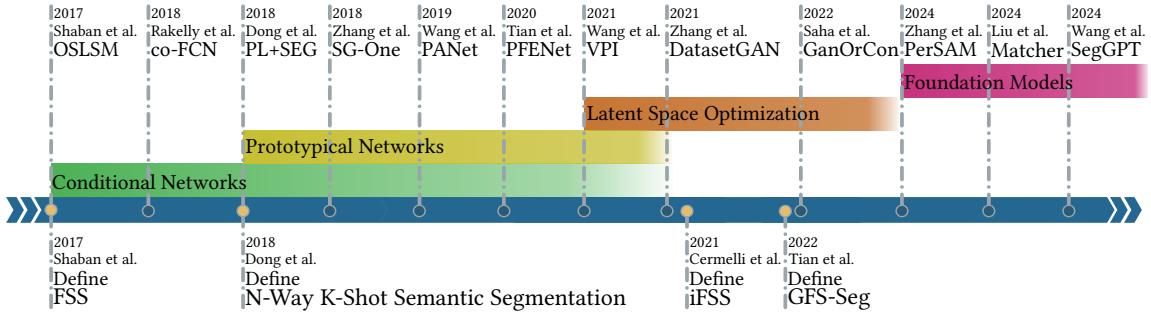


Fig. 2. Timeline illustrating the progression of the FSS research field. Some of the most consequential models are depicted at the top, with colored bands representing design trends and influences. Overlapping and fading bands suggest the sharing of concepts between multiple works to varying degrees. Below, key milestones mark the definition of the FSS problem and its variations.

as a function which takes as input the support set S and produces a set of parameters:

$$\theta = g(S)$$

On the other hand, the segmentation branch h extracts a feature volume from I_q using an embedding function ϕ . So let $F_q = \phi(I_q)$ be the feature volume, and F_q^{mn} be the feature vector at the spatial location (m, n) , the segmentation branch uses the parameters θ and the feature vector F_q^{mn} to predict the mask at the spatial location (m, n) as:

$$\hat{M}_q^{mn} = h(F_q^{mn}, \theta).$$

Rakelly et al. in [89] experimented with this setup by designing two different parameter sets θ . In the first experiment, θ is the feature volume extracted from the support set, which is fused with the query features in the segmentation branch. The resulting feature volume is then decoded to a binary mask \hat{M}_q by a small convolutional network f_θ that can be interpreted as a learned distance metric for retrieval from support to query. In the second experiment, θ is a set of linear classifier weights to be applied to the query feature volumes. The results of [89] show that the first approach leads to better outcomes, while Shaban et al. in [96] achieved slightly better predictions by using the parameter set θ to implement pixel-level logistic regression on the feature volume.

Lu et al. [78] proposed a similar overall architecture based on transformers [108] to predict θ . Similar to previous works in this category, θ is a set of weights for a classifier in charge of generating the final mask, and the feature extractor is pre-trained and shared between the conditioning and segmentation branches. Here, the key is the transformer within the segmentation branch as it utilizes the linear classifier weights θ from the conditioning branch as the query, and the feature vectors as the key and value. The resulting final weights are then used on a classifier to predict the subject mask on the query image.

Reckoning that different parameter sets θ can help the segmentation on different key aspects on the query image, Zhang et al. [135] propose leveraging the diversity of three distinct parameter sets. They propose the computation of three distinct similarity maps using these parameter sets, which are subsequently fused to generate a final prediction. The first parameter set, termed Peak Embedding, guides the segmentation process by identifying and emphasizing the most distinctive features in the query image. This is achieved by highlighting features with higher values relative to the subject embeddings, thereby directing attention towards salient image regions. The second parameter set, known as Global Embedding, focuses on capturing the overarching patterns of the subject within the support image. By computing

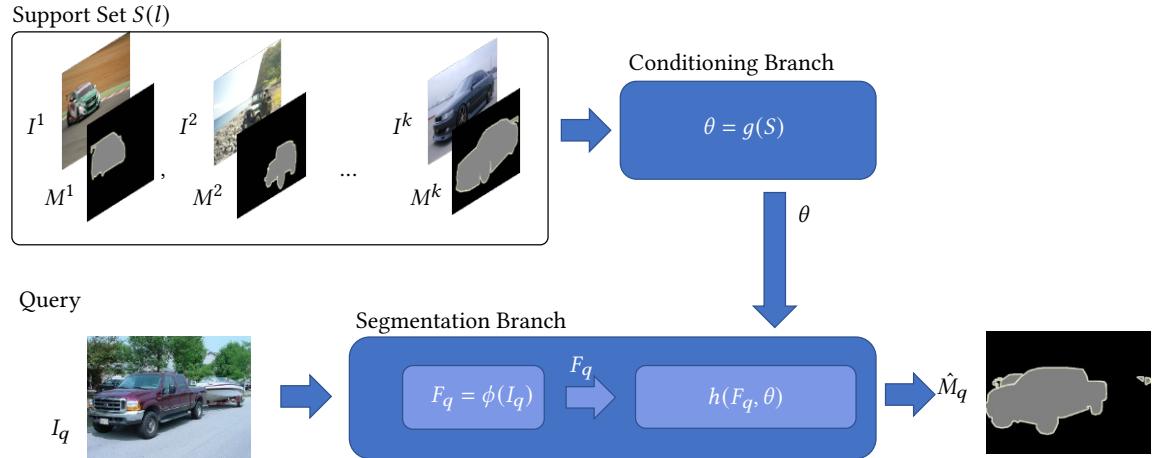


Fig. 3. Conditional networks adapted from [89]. Using the support set $S(l)$, the conditioning branch produces the parameter set θ . The segmentation branch then uses θ to predict a mask \hat{M}_q over the query image I_q .

the average vector of features within the object region, this embedding encapsulates the global context required for accurate segmentation. Lastly, the Adaptive Embedding is derived using an attention mechanism to discern the relative importance of object-related pixels via self-attention. This adaptive approach enables the model to dynamically adjust its focus based on the content and context of the input, enhancing segmentation performance in complex scenarios.

For these kinds of solutions, the choice of g and h is crucial, as the design of the conditioning branch and the interplay that its output θ has within the segmentation branch can significantly influence predictions. Specifically, the relation between the parameter set θ , the query features, and the expected prediction might be nontrivial and, therefore, challenging to learn for the segmentation branch. Conditional networks were the earliest model structure employed for the FSS task, and the shortcomings deriving from the use of two distinct functions for the support set and the query image have been mitigated with the introduction of the prototypical networks described in the next section.

3.2 Prototypical Networks

The challenges of FSS can be also approached with metric learning as reference framework. Prototypical Networks, which were initially introduced for **Few Shot Classification (FSC)** by Snell et al. [101], have been adapted for FSS by Dong et al. [24] and have become a popular strategy in many works in the field [64, 74, 113, 132, 136, 136]. Prototypical networks are based on the intuition that images can be represented by points in some embedding space where images of similar objects are represented by points close together, while points that are far away represent images of different objects. Therefore, computing the centroid of all the points associated with images of the same class in the support set will give you a point representing a prototype for that class. At inference time, the label of a new image can be predicted by looking at the nearest prototype. A graphical intuition of this concept is proposed in Figure 4. Given that segmentation can be modelled as pixel-wise classification, Dong et al. in [24] performs FSS by projecting each pixel of a query image in a learned feature space and then labelling each of them with the same label of the closest prototype as shown in Figure 5.

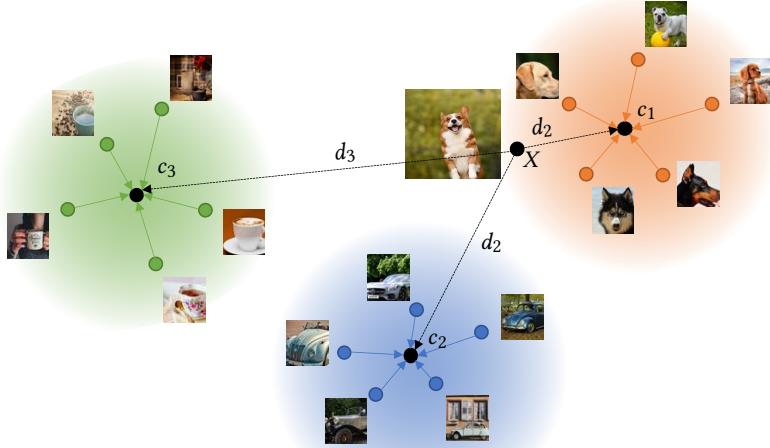


Fig. 4. Prototypical networks adapted from [101]. The prototypes are computed as the mean c_k of the embeddings from the same class in the support set identified in the picture with distinct colors. The label for the query embedding X is then assigned based on which prototype it is closer.

In this family of solutions, it is central to design representative prototypes and use reliable distance metrics that allow correct label assignments. This pivotal consideration hinges on two fundamental design choices: the computation of prototypes and selecting a suitable distance metric.

A commonly embraced approach for prototype generation is the application of *Masked Average Pooling* (MAP) [136]. MAP distils informative representations from the feature volume projected by support images through chosen feature extractors, typically ResNet50, ResNet101 [38], or VGG16 [100]. This process entails computing the Hadamard product—an element-wise multiplication—between the feature volume and the ground truth mask for each element in the support set. The averaged results of the Hadamard product yield a single feature vector embodying the class prototype. Practically, MAP aggregates feature vectors from spatial locations in the feature volume where the ground truth mask delineates the subject. While widely used backbones include ResNet50, ResNet101, and VGG16, a consensus on the preferred backbone has yet to emerge. Ongoing innovations in backbone design, such as combining multiple backbones [13] or designing new ones, hold promising for generating more informative embeddings, thereby potentially enhancing overall segmentation performance.

The second critical design choice involves selecting an appropriate distance metric to measure the similarity between prototypes and query features. Snell et al. advocate for Bregman distance divergences, such as the squared Euclidean distance. However, empirical findings by Wang et al. [113] suggest that a cosine distance metric offers greater stability and superior performance compared to Euclidean distance, arguing that the bounded nature of cosine similarity makes it more suitable for optimization, leading to its preference in subsequent works, including [74, 105, 113, 132, 136].

However, it is crucial to acknowledge certain limitations associated with prototype-based models in FSS. In applications where subjects can be viewed as compositions of multiple sub-parts, attempting to find a single prototype representing all parts may lead to suboptimal segmentation performances. For instance, a dog can be decomposed into a head, body, and limbs, each with distinct features. Acknowledging this limitation, Yang et al. [126] introduced

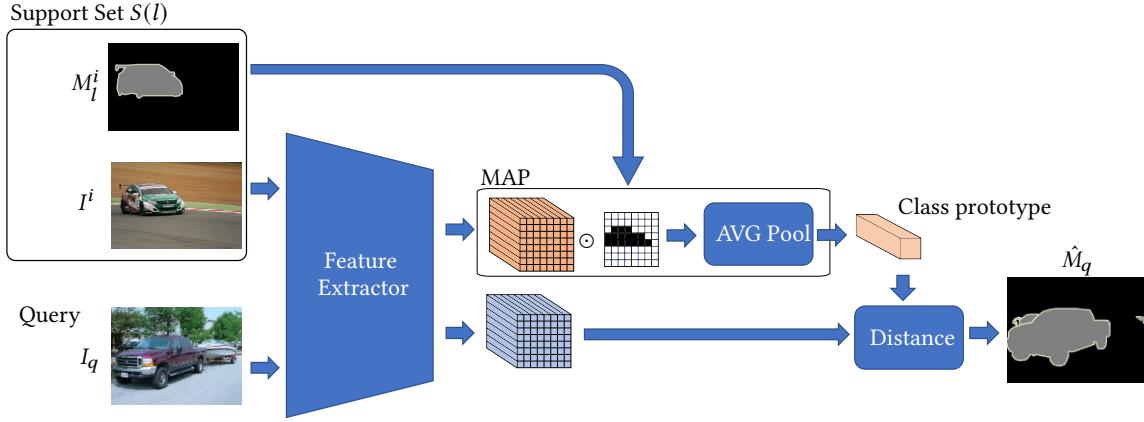


Fig. 5. Prototypical networks: a shared feature extractor gets a feature volume from both the support set and query images. The **Masked Average Pooling (MAP)** module takes the feature volume from the support set and masks its ground truth with the Hadamard product \odot to compute the class prototype. The prediction mask \hat{M}_q is calculated as a metric between the vector at each spatial location in the query feature volume with the class prototype.

a prototype mixture model, representing a class with more than one prototype to better capture the distinct macro feature of a subject class. Taking this concept further, Zhang et al. [131] and Wang et al. [111] extended the prototype strategy by adopting a graph-based solution. They leveraged a fully connected graph, where each node represents feature vectors from support and query images, and connection weights represent the similarities between two feature vectors. These connection weights were then employed for the pixel-level classification task.

In the pursuit of further enhancing the prototype strategy, another method with a number of prototypes greater than the number of classes was proposed by Li et al. [64]. They recognized that a prototypical model might get confused by similar classes present in the background of an image. To address this issue, they introduced a pseudo-class for areas of the image that correspond to high activation but are not labelled. This approach allows the model to exploit better the limited support set by inferring a new class.

To improve the separation of the prototypes, Okazawa [85] introduces a loss component to disincentive the cosine similarity between prototypes. In this way, the author aims to reduce the risk of computing prototypes that are too similar due to the sparse and low coverage of the latent space typical of the few-shot scenario. Conversely, Fan et al. [29] update the prototypes by selecting the query features that self-match the prototypes with high confidence. Such features, while representing high-confidence parts of the subject in the query, allow for the inclusion of features less represented in the query.

Another departure from prototypical networks is proposed by Wu et al. [119], recognizing that different classes may exhibit similar middle-level features despite having dissimilar appearances. To capitalize on this insight, they introduced a Meta-class Memory Module that aligns query and supports middle-level features to enhance the final segmentation mask prediction. The module learns meta-class embeddings, enabling effective utilization of similar features among different classes.

In summary, the design of prototypes and the choice of a suitable distance metric stand as the cornerstone of prototypical learning in the context of FSS. The utilization of MAP and the selection of an appropriate distance metric play pivotal roles in achieving robust and accurate segmentation models. However, it is imperative to acknowledge

certain limitations associated with prototype-based models in scenarios where subjects exhibit compositional complexity. In these scenarios, innovative solutions such as prototype mixture models or graph-based approaches are needed. As we delve into the challenges of FSS, a crucial aspect comes into focus: the exploration of optimal backbones for feature extraction. The selection of a backbone, be it a well-established one like ResNet50, ResNet101, or VGG16, or the innovation of new backbones, has a pivotal role in defining the feature space and influencing the quality of the extracted embeddings. This choice becomes a determining factor shaping the overall performance of the downstream segmentation task. Hence, delving into the possibilities of diverse backbones becomes not only a worthwhile endeavour but a promising avenue to unlock improved segmentation performance.

3.3 Latent space optimization

The effectiveness of FSS model designs can often be related to their representation of classes and thus in the corresponding feature space. This section presents various approaches for latent space class representation in FSS. Firstly, we discuss Generative Adversarial Networks (GANs) and how their regular latent space can be exploited. Secondly, we introduce studies that have achieved results similar to those of GANs but using simpler networks based on contrastive learning. Finally, we explore works that model the variability of subject classes using probabilistic frameworks such as Variational Autoencoders (VAEs).

3.3.1 Generative models. In certain applications of object segmentation, it is possible to have plenty of data available, but not enough label information; this scenario is particularly common in tasks like part segmentation. For instance, while it is possible to access to numerous images of objects like faces or cars through public datasets, creating detailed masks for their individual parts often requires considerable effort due to the absence of pre-existing parts labels. To fully exploit already available datasets and minimize the need for new labels, some works [95, 106, 137] have resorted to the framework of FSS and developed methods based on generative models. The core idea is that good performing GANs such as StyleGAN [52] or StyleGAN2 [53] have an internal representation of images that effectively synthesize both semantic and geometrical information and disentangle the latent factors of variation. This intuition has been supported by Karras et al. in [52] where the authors show that interpolating latent codes of different images produces a synthetic image that is still qualitatively good. Given these intuitions, the internal representation of a GAN generated image should be more informative than the feature extracted by other means and thus lead to better segmentation performance.

With this reasoning, Zhang et al. in [137] and Tritrong et al. in [106] exploit the regularity of StyleGAN [52] and StyleGAN2 [53] latent spaces to predict part segmentation masks. In their models, the segmentation is performed by a FSS module which predicts the masks from the GAN’s internal representation of a generated image. The practical implementation of this design involves a sequential process. Firstly, in the presence of abundant unlabeled target domain images, the GAN is trained on this data to optimize its latent space. Subsequently, the training set for the FSS module is constructed by sampling k generated images along with their corresponding internal representations. The k images are then manually annotated by a human annotator, serving as the ground truth for training the FSS module. Once the entire model is trained, Tritrong in [106] proposes to perform latent code optimization to make the GAN generate an image as close as possible to the test image and record its internal representation. The generated image and its internal representation are finally fed to the segmentation module to predict the parts segmentation mask. A depiction of the inference procedure for this model is proposed in Figure 6.

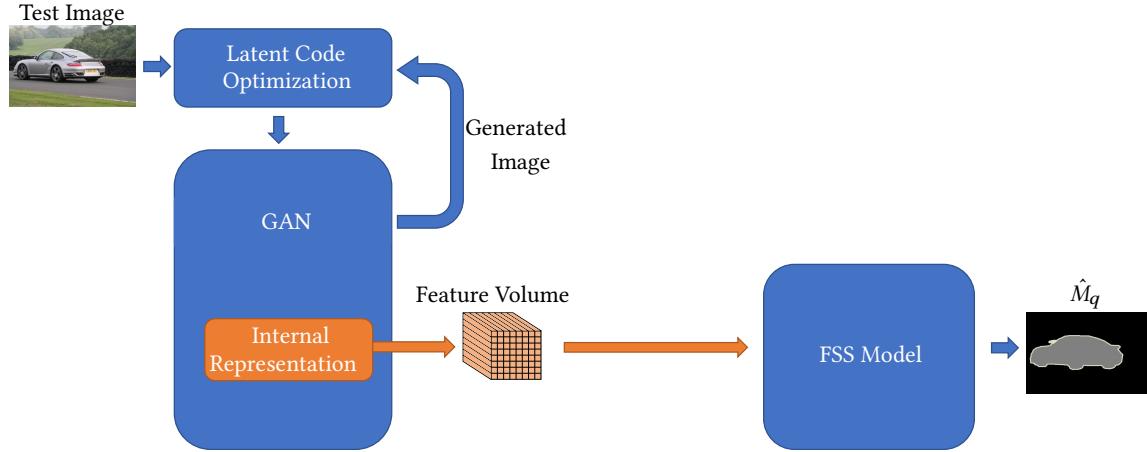


Fig. 6. Exploitation of **GAN** regular latent representation for **FSS**. A **GAN** is employed to generate an image that closely resembles the test image by latent code optimization. Once the generated image is sufficiently similar to the test image, its internal representation is extracted and fed into a **FSS** module to predict the final segmentation mask \hat{M}_q .

As noted by the authors themselves, the latent code optimization step is computationally expensive. Additionally, the results of their experiments also suggest that the generated image might still be dissimilar to the test image, leading to poor performance. Tritrong et al. in [106] and Zhang et al. in [137] propose to overcome such limitations by using the trained model to generate a large synthetic dataset as shown in Figure 7. The process involves sampling random new images and predicting the parts segmentation masks via the **FSS** module as before. The new synthetic dataset, composed of the generated images and inferred masks, can later be used to train any off-the-shelf segmentation model. Tritrong et al. in [106] validate this approach by showing that generating a new parts segmentation dataset avoiding the latent code optimization step reduces inference time and improves segmentation performance.

GANs can have complex training procedures, but once trained, they can be used as feature extractors as in [95, 106], or to produce new synthetic data to mitigate the data scarcity regime such as in [137]. The wide range of strategies in this family of approaches branch makes it hard to define a general training procedure as it is possible for conditional networks and prototypical networks where episodic learning described in subsection 2.1 is standard practice. Moreover, GAN-based FSS models are not restricted to the conventional values of 1 or 5 shots used in conditional and prototypical networks. For instance, Zang et al. in [137] generate 20 images from the GAN to train an MLP **FSS** module, similarly, Tritrong et al. in [106] experiments with 1, 5 and 10 images to train the **FSS** module, while Saha et al. in [95] uses 20 shots.

In conclusion, while GANs have shown promise for **FSS**, particularly in scenarios with abundant unlabeled data, their complex and computationally intensive training process raises questions about alternative approaches. The benefits of GANs' internal representation and regularity come with challenges, notably in the latent code optimization step. Section 3.3.2 delves into works exploring simpler architectures focusing on achieving comparable results in **FSS** without the complexities associated with training GANs.

3.3.2 Contrastive learning. In the realm of **FSS**, the effectiveness of **GANs** as powerful tools for generating regular internal representations has been acknowledged. However, Saha et al. in [95] highlight how **GANs** can be challenging and time-consuming to train and their performance may decrease when using latent space optimization to get a test

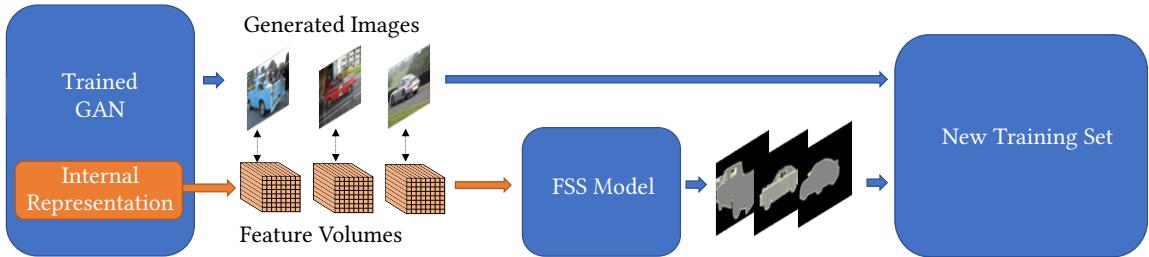


Fig. 7. A generative model is used to create a synthetic dataset with a limited annotation effort. First a **GAN** is trained on the target dataset. A small number of synthetic images are labelled and in couple with their internal representations, are used to train a **FSS** module. During inference, an unlimited number of images and internal representations can be generated from the **GAN** and fed into the **FSS** module to produce masks. The new synthetic dataset formed by the generated images and their masks can be used to train any off-the-shelf segmentation model.

image latent code. To address these issues, Saha et al. introduced the use of contrastive learning for the feature extractor, demonstrating that this technique can make the backbone consistently produce latent representations that outperform **GAN**-based models, without increasing its complexity.

Contrastive learning, as described in [4, 19, 25], involves training a model on a self-supervised proxy task to make its internal representations invariant to augmentations. This is achieved by constructing a dataset of image pairs, where one image is taken from the original dataset and the other is either a transformed version of the same image or a different one. The model is then trained to classify whether the two images represent the same subject or not. Once the model is trained it can be used as a feature extractor by feeding it with the test image and recording its internal representation. Saha et al. use the feature volume extracted in this manner to predict the part segmentation masks using a U-Net [94] inspired network. This model achieves slightly better results than the **GAN** based model proposed by Zhang et al. [137] while being easier to train.

In conclusion, contrastive learning emerges as a versatile tool for regularization in feature spaces of models employing feature extractors. This technique, applicable to various design strategies in **FSS**, holds the potential to enhance performance without escalating model complexity, prompting further exploration in the field.

3.3.3 Variational Auto Encoders. Within the domain of **FSS**, the prevalence of prototypical networks has been a cornerstone for representation in multiple works. However, the deterministic nature of their feature extractor presents limitations, especially in dealing with intra-class variations and noise in one-shot scenarios. Additionally, compressing an entire class into a single deterministic prototype vector can result in the loss of important structural information of the objects in the scene. Aligning with the pursuit of a more regular latent representation Wang et al. [110] propose an innovative approach based on Variational Autoencoders (VAEs) [55, 93]. Departing from traditional prototypical learning, this model introduces a shift toward a probabilistic framework, learning a probability distribution of prototypes for a given class. This allows for a more comprehensive encoding of object variations. In a later work [103], Sun et al. further enhanced the model by incorporating a variational attention mechanism to highlight the foreground of the test image, resulting in improved overall segmentation performance.

3.4 Foundation Models

Recent developments in the **AI** community have seen the emergence of Foundation Models, which are versatile models trained on extensive datasets and capable of adapting to diverse downstream tasks [6]. Notably, while prominent

examples such as BERT [54] and GPT [9, 88] originate from the **Natural Language Processing (NLP)** community, the realm of **Computer Vision (CV)** also boasts its own foundation models like CLIP [87] and SAM [56].

CLIP can be framed as a bridge between **NLP** and **CV** by establishing a unified embedding space where images and text are represented. Through contrastive learning, CLIP aligns image and text embeddings, facilitating cross-modal understanding and enabling tasks such as classification, retrieval, and generation. On the other hand, Kirillov et al. developed Segment Anything (SAM), a segmentation foundation model based on a novel promptable segmentation framework. Thanks to a data engine collecting 11M image-mask pairs, SAM is trained on vast and diverse datasets, allowing it to segment any objects in visual contexts. SAM accepts handcrafted prompts and returns the expected mask, supporting both sparse (points, boxes, text) and dense (masks) prompts. SAM stands out as a generalist vision foundation model for image segmentation, offering flexibility through a diverse range of input prompts. However, it lacks the ability to recognize the type of each segmented object.

The release of vision foundation models has sparked excitement in the **CV** community, including in the **FSS** research stream. Foundation models present a novel opportunity for **FSS**, as they can be leveraged to segment classes for which tailored datasets are not available. Notably, strategies for utilizing vision foundation models in **FSS** can be drastically different from the one already presented in this paper so far and require a separate discussion. In addition, in this category, we can identify three major types: prompt engineering, multi-modal approaches, and generalist models. This section delves deeper into these strategies, highlighting their significance in advancing the field of computer vision.

3.4.1 Prompt engineering. In the realm of image segmentation, foundation models like SAM offer remarkable capabilities but rely on additional guidance in the form of prompts tailored to specific object classes. SAM, designed to accept prompts in various forms such as point boxes, text, and masks, has spurred research efforts aimed at enhancing its specialization through prompt engineering.

One notable approach in this domain is introduced by Liu et al. with Matcher [75], a model designed for one-shot segmentation, leveraging SAM as its foundational model. Matcher begins by embedding both the support and query images (in this context called reference and target image) using the DINOv2 feature extractor [86]. Given the correspondence between spatial locations in the feature volume and patches from the images, Matcher computes a similarity matrix between features to identify similar patches in the images. Points within patches achieving high similarity scores are selected as prompts for SAM, whose predictions are aggregated to generate the final segmentation output.

Similarly, Zhang et al. propose PerSAM [134], a training-free personalization approach for SAM. PerSAM customizes SAM using one-shot data, comprising a reference image and a rough mask of the desired concept. Initially, PerSAM generates a location confidence map for the target object in the test image based on feature similarities. Subsequently, it selects two points as positive-negative location priors based on confidence scores, encoding them as prompt tokens for SAM's decoder to perform segmentation.

Both Matcher and PerSAM heavily rely on the feature volumes of reference and test images to generate appropriate prompts, emphasizing the importance of flexible and informative embeddings, drawing thus similarities with prototypical learning discussed in previous subsection. Additionally, SAM's flexibility in accommodating various prompt types suggests the potential for further exploration, particularly regarding the utilization of other available prompt types. In the subsequent subsection, we delve into works that leverage text prompts as part of a multimodal approach.

3.4.2 Multimodal. Multimodal approaches harness the synergy between different data modalities to enhance understanding and performance. One prominent application is the alignment of visual and textual embeddings, a concept

dating back to proposals as far as 2014 [27, 51, 57, 91]. These techniques facilitate correlating subjects in images with their textual descriptions, laying the groundwork for text-promptable segmentation models. These models, not requiring any specific training for a novel class fall closely to FSS, albeit requiring a text description instead of a segmentation example.

The introduction of CLIP reinvigorated research in this field. CLIP’s training on a large and diverse datasets makes it highly flexible, enabling zero-shot classification of images. This flexibility has inspired many authors to propose one-shot segmentation models where the prompt is a textual description of the subject. Numerous works in this field [79, 117, 120, 140] have adopted an encoder-decoder architecture. Typically, the encoder leverages a pre-trained CLIP to extract aligned visual and text embeddings. The alignment of these embeddings, originating from inputs of different natures, empowers the decoder module to predict segmentation masks by focusing its design on the inner product between the text embeddings and the visual embeddings. While multimodal models utilizing text prompts alleviate the need for extensive segmentation datasets, it’s worth noting that for certain application-specific domains, describing targets as texts may pose challenges. For example textual ambiguity could render this approach less than ideal in such cases.

In conclusion, while the concept of aligning textual and visual embeddings is not new, recent advancements, particularly with models like CLIP, have propelled this research area forward. One-shot segmentation models driven by textual prompts hold promise for various applications, albeit with considerations for domain-specific challenges and ambiguities in textual descriptions.

3.4.3 Generalist models. Generalist models, often built upon foundation models, exhibit remarkable adaptability across diverse applications. Leveraging techniques like In-Context Learning [1, 8, 36], primarily developed within NLP, these models can swiftly accommodate new tasks with minimal examples or prompts. This adaptability extends to FSL techniques, enabling their application in FSS tasks. In-context learning, pioneered in NLP, utilizes language sequences as a versatile interface, facilitating rapid adaptation to various language-centric tasks with minimal examples. However, while In-Context Learning has flourished in NLP, its exploration in computer vision remains nascent due to significant task disparities. Unlike NLP tasks, which revolve around discrete language tokens, CV tasks encompass diverse output representations, posing challenges in defining task prompts. Addressing this disparity, Wang et al.[114, 115] introduced an approach where images serve as the interface for visual perception. Their models, Painter [114] and its successor SegGPT [115], adopt the principles of NLP models, utilizing images both as prompts and outputs. By framing dense prediction tasks as image inpainting, these models leverage three-component tensor inputs: an image paired with its label (akin to a support set in FSS literature), and a target image for prediction. Similarly, Meng et al. proposed SEGIC [80], a model combining foundation models and in-context learning. It employs a pre-trained foundation model for feature extraction, followed by a novel “correspondence discovery” step. This step establishes semantic and geometric correspondences between the in-context and target feature maps, providing crucial guidance for segmentation. Subsequently, “in-context instructions” are extracted based on the in-context samples and these correspondences. Finally, a lightweight decoder leverages this information to generate the segmentation mask for the target image. A key feature of these models and training procedures is their ability to leverage larger pre-training datasets. To this end, to guarantee SegGPT its flexibility, a key factor is its pre-training on a large combination of existing datasets, including ADE20K, MS-COCO [66], PASCAL VOC [26], Cityscapes [21], LIP [65], PACO [90], CHASE_DB1 [30], DRIVE [102], HRF [10], STARE [42], iSAID [118], and loveDA [112].

While foundational and generalist models hold promise for vision tasks, including FSS, further investigation is imperative. Their superiority over task-specific models, particularly in diverse domains, necessitates comprehensive validation. Moreover, challenges such as large and diverse training sets and training procedures, as well as domain-specific performance variations and parameter tuning complexities, warrant attention. Exploring alternatives like backbones based on space-state models, such as MAMBA [33] and VisionMAMBA[142], could offer simplified architectures without compromising performance, paving the way for more efficient vision models.

4 MODELS EVALUATION

To train, test, and evaluate the variety of FSS models presented in previous sections, it is crucial to have publicly available standard datasets with densely annotated images that allow episodic training. Furthermore, a common evaluation metric is essential to rank works in the literature. In this section, we first present the metrics used to evaluate segmentation models, followed by a description of the publicly available benchmark datasets for FSS. In addition, we present a review of results compiled from the papers of some of the most significant models in the literature, showcasing their performance on the presented common standard benchmarks. Finally, in subsection 4.4 we present a selection of predictions obtained under the same conditions to qualitatively compare predictions of different models in multiple backbone and training dataset combinations.

4.1 Metrics

As common in segmentation tasks, the evaluation metrics commonly found in the FSS literature are based on the **Intersection over Union (IoU)**. Two widely used metrics in FSS are the **mean Intersection over Union (mIoU)** or the **Foreground Background Intersection over Union (FB-IoU)**. Recalling that for the task of segmenting a subject of class c as pixel-wise classification the **IoU** is defined as:

$$IoU_c = \frac{TP_c}{TP_c + FP_c + FN_c} \quad (2)$$

where TP_c , FP_c and FN_c are the number of pixels that are true positive, false positive and false negative respectively. The **mIoU** is defined as the average of **IoU** for all classes:

$$mIoU = \frac{1}{C} \sum_{c=1}^{l=C} IoU_c \quad (3)$$

where C is the number of classes.

Another common metric found in FSS is **Foreground Background Intersection over Union (FB-IoU)**. The FB-IoU is a variation of the **mIoU** as it ignores categories by only considering foreground and background ($C = 2$). However, this approach has been criticized for its limitations. In particular, Zhang et al. [132] argue that ignoring categories can lead to biased results when there is a class imbalance in the dataset, such as in MS COCO or PASCAL VOC. The **FB-IoU** may also not accurately reflect the performance of a model, as failing to segment small objects will not necessarily be reflected in **FB-IoU** scores. The reason is that if the background occupies a large portion of the scene, a small object will have little influence over the **FB-IoU**.

4.2 Public Datasets

While introducing the problem of **Few Shot Semantic Segmentation**, Shaban et al. in [96] introduced a new dataset to simulate segmentation episodes based on the data of the original PASCAL VOC 2012 [26] dataset and its extended

annotations in SDS [37] naming it PASCAL-5ⁱ. Given the set L of all 20 subject classes from the PASCAL VOC 2012 dataset Shaban et al. in [96] first sample 5 classes to create the label set L_{test} and uses the remaining 15 to build L_{train} . Sampling different classes for L_{test} and L_{train} allow to create 4 folds from the original dataset as we can build $L_{test} = \{4i + 1, \dots, 4i + 5\}, i \in [0, 3]$. Once L_{test} and L_{train} are defined, Shaban et al. in [96] form the train dataset D_{train} by picking all the $(image, mask)$ pairs from the PASCAL VOC train subset that contains subject from L_{train} . Similarly, the test dataset D_{test} contains all $(image, mask)$ pairs in the validation subset of PASCAL VOC containing subjects from L_{test} . The two dataset D_{train} and D_{test} are then used to sample training and testing episodes as explained in subsection 2.1.

The generalization capacity of a model is a crucial property in FSS, therefore, the few classes of PASCAL-5ⁱ can make it unideal as a benchmark. To this end, Nguyen et al. in [84] introduced a more challenging dataset named COCO-20ⁱ where episodes are created similarly to PASCAL-5ⁱ, but now with 80 classes based on MS COCO [66]. The 80 classes are sampled with the same procedure as for PASCAL-5ⁱ leading to 4 possible folds of 20 class each from the original dataset. While COCO-20ⁱ has more classes than PASCAL-5ⁱ Li et al. [63] still found it inadequate to test FSS models as a benchmark dataset for this task should have a large number of classes while requiring only few samples for class. To this end, they introduced a totally new dataset called FSS-1000 designed specifically for FSS, with 1000 classes having 10 images per class. Unlike COCO-20ⁱ and PASCAL-5ⁱ, this dataset only segments one class per image. This means that even if multiple class subjects are present in a single image, only one will be segmented, making it possible to evaluate only FB-IoU metric.

4.3 Benchmark results

In this section, we present a comprehensive collection of benchmark results from various seminal works in the FSS field. By synthesizing these findings from the literature, we provide the reader with a consolidated resource, complete with details on the year and venue of each model's presentation. This presentation format facilitates easy comparison between different works and enables the observation of performance trends across recent years. Moreover, the tables presented herein offer insights into the attention received by different benchmark datasets, highlighting those that are more extensively explored and those that are less so. For instance, it becomes evident that while the older benchmark dataset PASCAL-5ⁱ has been extensively tested, COCO-20ⁱ has rapidly gained popularity, and FSS-1000 remains less prominent in the literature.

In reporting results for specialist FSS models, we provide metrics for both 1 Shot and 5 Shot settings, along with details on the backbone architecture utilized. Specifically, for benchmark datasets such as Pascal-5ⁱ and COCO-20ⁱ we report the mIoU in Table 1 and Table 3 and the FB-IoU in Table 2 and Table 4, while for FSS-1000, we report the FB-IoU in Table 5. In table Table 6, we report the scores of generalist models from the literature and compare them to some of the more modern specialist ones.

As is customary in the literature of FSS for benchmark datasets where mIoU is employed as a performance metric, we report the mIoU on each fold as well as its average over the 4 folds. Conversely, folds are typically not used when evaluating performance with the FB-IoU metric. As a result, we only report the performance metric on the entire dataset in such cases.

In essence, these tables serve not only as repositories of benchmark results but also enhance the value of the paper by providing a centralized and comparative analysis of performance across different models and datasets. This comprehensive analysis aids researchers in assessing the state-of-the-art and identifying promising avenues for future exploration.

Table 1. Evaluation of 1-shot and 5-shot semantic segmentation specialist models on PASCAL-5ⁱ using mean-IoU scores available in the literature. We report the average over the four folds, and, when available, the results on each individual fold. Results for PPNet without using additional unlabeled data are reported.

Backbone	Method	1 Shot					5 Shot				
		Fold-0	Fold-1	Fold-2	Fold-3	MIoU%	Fold-0	Fold-1	Fold-2	Fold-3	MIoU%
VGG16	OSLSM [96] (BMVC 2017)	33.6	55.3	40.9	33.5	40.8	35.9	58.1	42.7	39.1	43.9
	PL + SEG [24] (BMVC 2018)	-	-	-	-	61.2	-	-	-	-	62.3
	co-FCN 2018 [89] (ICLR 2018)	36.7	50.6	44.9	32.4	41.1	37.5	50.0	44.1	33.9	41.4
	SG-One [136] (TCYB 2019)	40.2	58.4	48.4	38.4	46.3	41.9	58.6	48.6	39.4	47.1
	PANet [113] (ICCV 2019)	42.3	58.0	51.1	41.2	48.1	51.8	64.6	59.8	46.5	55.7
	AMP [98] (ICCV 2019)	41.9	50.2	46.7	34.7	43.4	41.8	55.5	50.3	39.9	46.9
	FWB [84] (ICCV 2019)	47.0	59.6	52.6	48.3	51.9	50.9	62.9	56.5	50.1	55.1
	CRNet [72] (CVPR 2020)	-	-	-	-	55.2	-	-	-	-	58.5
	FSS-1000 [63] (CVPR 2020)	-	-	-	-	-	37.4	60.9	46.6	42.2	56.8
	PFENet [105] (TPAMI 2020)	56.9	68.2	54.4	52.4	58.0	59.0	69.1	54.8	52.9	59.0
	RPMM [74] (ECCV 2020)	47.1	65.8	50.6	48.5	53.0	50.0	66.5	54.9	47.6	54.0
	SST [141] (IJCAI 2020)	50.9	63.0	53.6	49.6	54.3	52.5	64.8	59.5	51.3	57.0
	HSNet [81] (ICCV 2021)	59.6	65.7	59.6	54.0	59.7	64.9	69.0	64.1	58.6	64.1
	MM-Net [119] (ICCV 2021)	57.1	67.2	56.6	52.3	58.3	56.6	66.7	53.6	56.5	58.3
	BAM [58] (CVPR 2022)	63.2	70.8	66.1	57.5	64.4	67.4	73.1	70.6	64.0	68.8
	FECANet [68] (TMM 2023)	66.5	68.9	63.6	58.3	64.3	68.6	70.8	66.7	60.7	66.7
ResNet50	CANet [132] (ICCV 2019)	52.5	65.9	51.3	51.9	55.4	55.5	67.8	51.9	53.2	57.1
	PGNet [131] (ICCV 2019)	56.0	66.9	50.6	50.4	56.0	57.7	68.7	52.9	54.6	58.5
	CRNet [72] (CVPR 2020)	-	-	-	-	55.7	-	-	-	-	58.8
	PMMs [126] (ECCV 2020)	52.0	67.5	51.5	49.8	55.2	55.0	68.2	52.9	51.1	56.8
	PPNet [74] (ECCV 2020)	47.8	58.8	53.8	45.6	51.5	58.4	67.8	64.9	56.7	62.0
	RPMMs [126] (ECCV 2020)	55.2	66.9	52.6	50.7	56.3	56.3	67.3	54.5	51.0	57.3
	PFENet [105] (TPAMI 2020)	61.7	69.5	55.4	56.3	60.8	63.1	70.7	55.8	57.9	61.9
	SST [141] (IJCAI 2020)	54.4	66.4	57.1	52.5	57.6	58.6	68.7	63.1	55.3	61.4
	SimPropNet [31] (IJCAI 2020)	54.9	67.3	54.5	52.0	57.2	57.2	68.5	58.4	56.1	60.0
	LTM [128] (MMM 2020)	52.8	69.6	53.2	52.3	57.0	57.9	69.9	56.9	57.5	60.6
	ASGNet [62] (CVPR 2021)	58.8	67.9	56.8	53.7	59.3	63.7	70.6	64.2	57.4	63.9
	RePRI [7] (CVPR 2021)	60.2	67.0	61.7	47.5	59.1	64.5	70.8	71.7	60.3	66.8
	CWT [78] (ICCV 2021)	56.3	62.0	59.9	47.2	56.4	61.3	68.5	68.5	56.6	63.7
	MM-Net [119] (ICCV 2021)	62.7	70.2	57.3	57.0	61.8	62.2	71.5	57.5	62.4	63.4
	HSNet [81] (ICCV 2021)	64.3	70.7	60.3	60.5	64.0	70.3	73.2	67.4	67.1	69.5
	VAT [41] (ECCV 2022)	67.6	71.2	62.3	60.1	65.3	72.4	73.6	68.6	65.7	70.0
	BAM [58] (CVPR 2022)	69.0	73.6	67.6	61.1	67.8	70.6	75.1	70.8	67.2	70.9
	DCAMA [97] (ECCV 2022)	67.5	72.3	59.6	59.0	64.6	70.5	73.9	63.7	65.8	68.5
	IPMT [73] (2022)	72.8	73.7	59.2	61.6	66.8	73.1	74.7	61.6	63.4	68.2
	FECANet [68] (TMM 2023)	69.2	72.3	62.4	65.7	67.4	72.9	74.0	65.2	67.8	70.0
	ABCNet [45] (CVPR2023)	68.8	73.4	62.3	59.5	66.0	71.7	74.2	65.4	67.0	69.6
	PMNet [16] (WACV 2024)	67.3	72.0	62.4	59.9	65.4	73.6	74.6	69.9	67.2	71.3
ResNet101	FWB [84] (ICCV 2019)	51.3	64.5	56.7	52.2	56.2	54.8	67.4	62.2	55.3	59.9
	PPNet [74] (ECCV 2020)	52.7	62.8	57.4	47.7	55.2	60.3	70.0	69.4	60.7	65.1
	DAN [111] (ECCV 2020)	54.7	68.6	57.8	51.6	58.2	57.9	69.0	60.1	54.9	60.5
	PFENet [105] (TPAMI 2020)	60.5	69.4	54.4	55.9	60.1	62.8	70.4	54.9	57.6	61.4
	VPI [110] (WACV 2021)	53.4	65.6	57.3	52.9	57.3	55.8	67.5	62.6	55.7	60.4
	RePRI [7] (CVPR 2021)	59.6	68.6	62.2	47.2	59.4	66.2	71.4	67.0	57.7	65.6
	HSNet [81] (ICCV 2021)	67.3	72.3	62.0	63.1	66.2	71.8	74.4	67.0	68.3	70.4
	CWT [78] (ICCV 2021)	56.9	65.2	61.2	48.8	58.0	62.6	70.2	68.8	57.2	64.7
	CyCTR [133] (NIPs 2021)	69.3	72.7	56.5	58.6	64.3	73.5	74.0	65.2	66.6	66.6
	ASGNet [62] (CVPR 2021)	59.8	67.4	55.6	54.4	59.3	64.6	71.3	64.2	57.3	64.4
	VAT [41] (ECCV 2022)	68.4	72.5	64.8	64.2	67.5	73.3	75.2	68.4	69.5	71.6
	DCAMA [97] (ECCV 2022)	65.4	71.4	63.2	58.3	64.6	70.7	73.7	66.8	61.9	68.3
Swin-B	IPMT [73] (NeurIPS 2022)	71.6	73.5	58.0	61.2	66.1	75.3	76.9	59.6	65.1	69.2
	API [103] (Pattern Recognition 2023)	53.4	65.6	57.3	52.9	57.3	55.8	67.5	62.6	55.7	60.4
	ABCNet [45] (CVPR 2023)	65.3	72.9	65.0	59.3	65.6	71.4	75.0	68.2	63.1	69.4
	PMNet [16] (WACV 2024)	71.3	72.4	66.9	61.9	68.1	74.9	75.5	75.3	69.8	73.9

Manuscript submitted to ACM

Table 2. Evaluation of 1-shot and 5-shot semantic segmentation specialist models on PASCAL-5ⁱ using FB-IoU scores available in the literature.

Backbone	Method	1 Shot	5 Shot
VGG16	OSLSM [96] (BMVC 2017)	61.3	61.5
	co-FCN [89] (ICLR 2018)	60.1	60.2
	PL + SEG [24] (BMVC 2018)	61.2	62.3
	SG-One [136] (TCYB 2019)	63.9	65.9
	PANet [113] (ICCV 2019)	66.5	70.7
	PFENet [105] (TPAMI 2020)	72	72.3
	HSNet [81] (ICCV 2021)	73.4	76.6
	BAM [58] (CVPR 2022)	77.26	81.1
	MMnet [119] (ICCV 2021)	78.01	80.5
	APANet [17] (TMM 2022)	71.3	75.2
	AMP [99] (ICCV 2019)	62.2	63.8
	FECANet [68] (TMM 2023)	76.2	77.6
ResNet50	CANet [132] (ICCV 2019)	66.2	69.6
	PGNet [131] (ICCV 2019)	69.9	70.5
	PPNet [74] (ECCV 2020)	69.19	75.76
	PFENet [105] (TPAMI 2020)	73.3	73.9
	HSNet [81] (ICCV 2021)	76.7	80.6
	BAM [58] (CVPR 2022)	79.71	82.18
	VAT [41] (ECCV 2022)	77.4	80.9
	MMnet [119] (ICCV 2021)	80.44	83.23
	SAGNN [121] (CVPR 2021)	73.2	73.3
	SCLNet [129] (CVPR 2021)	71.9	72.8
	APANet [17] (TMM 2022)	72.3	77.4
	ASGNet [62] (CVPR 2021)	69.2	74.2
	CMN [122] (ICCV 2021)	72.3	72.8
	MANet [2] (arXiv 2021)	71.4	75.2
	DCP [59] (IJCAI 2022)	75.6	79.7
	MFGN [125] (MDPI 2022)	76.4	79.3
	CRNet [72] (CVPR 2020)	66.8	71.5
	CMNet [71] (TMM 2022)	73.5	74.1
	SimPropNet [31] (IJCAI 2020)	73	72.9
	CMN [123] (ICCV 2021)	72.3	72.5
ResNet101	ASNet [50] (CVPR 2022)	77.7	80.4
	SCL(CANet) [130] (CVPR 2021)	70.3	70.7
	SCL(PFENet) [130] (CVPR 2021)	71.9	72.8
	IPMT [73] (NeurIPS 2022)	77.1	81.4
	FECANet [68] (TMM 2023)	78.7	80.7
	A-MCG [43] (AAAI 2019)	61.2	62.2
	PFENet [105] (TPAMI 2020)	72.9	73.5
	PPNet [74] (ECCV 2020)	70.9	77.5
	DAN [111] (ECCV 2020)	71.9	72.3
	HSNet [81] (ICCV 2021)	77.6	80.6
	CyCTR [133] (NIPs 2021)	72.9	75
Swin-B	VAT [41] (ECCV 2022)	78.8	82
	APANet [17] (TMM 2022)	74.9	78.8
	MFGN [125] (MDPI 2022)	76.8	79.6
	ASGNet [62] (CVPR 2021)	71.7	75.2
	IPMT [73] (NeurIPS 2022)	78.5	80.3
Swin-B	HSNet [81] (ICCV 2021)	77.9	81.2
	DCAMA [97] (ECCV 2022)	78.5	82.9

Table 3. Evaluation of 1-shot and 5-shot semantic segmentation specialist models on MS COCO-20ⁱ using mean-IoU scores available in the literature. We report the average over the four folds and, when available, the results for each individual fold. Results for PPNet without using additional unlabeled data are reported.

Backbone	Method	1 Shot					5 Shot				
		Fold-0	Fold-1	Fold-2	Fold-3	MIoU%	Fold-0	Fold-1	Fold-2	Fold-3	MIoU%
VGG16	PANet2019 [113] (ICCV 2019)	28.7	21.2	19.1	14.8	20.9	39.4	28.3	28.2	22.7	29.7
	FWB [84] (ICCV 2019)	18.4	16.7	19.6	25.4	20.0	20.9	19.2	21.9	28.4	22.6
	PFENet [105] (TPAMI 2020)	35.4	38.1	36.8	34.7	36.3	38.2	42.5	41.8	38.9	40.4
	BAM [58] (CVPR 2022)	39.0	47.0	46.4	41.6	43.5	47.0	52.6	48.6	49.1	49.3
ResNet50	PANet [113] (ICCV 2019)	31.5	22.6	21.5	16.2	23.0	45.9	29.2	30.6	29.6	33.8
	PMMs [126] (ECCV 2020)	29.3	34.8	27.1	27.3	29.6	33.0	40.6	30.3	33.3	34.3
	BriNet [127] (BMVC2020)	32.9	36.2	37.4	30.9	34.4	-	-	-	-	-
	RPMMs [126] (ECCV 2020)	29.5	36.8	29.0	27.0	30.6	33.8	42.0	33.0	33.3	35.5
	PPNet [74] (ECCV 2020)	34.5	25.4	24.3	18.6	25.7	48.3	30.9	35.7	0.3	36.2
	HFA [67] (TIP 2021)	28.7	36.0	30.2	33.3	32.0	32.7	42.1	30.4	36.2	35.3
	MM-Net [119] (ICCV 2021)	34.9	41.0	37.2	37.0	37.5	37.0	40.3	39.3	36.0	38.2
	RePRI [7] (CVPR 2021)	31.2	38.1	33.3	33.0	34.0	38.5	46.2	40.0	43.6	42.1
	HSNet [81] (ICCV 2021)	36.3	43.1	38.7	38.7	39.2	43.3	51.3	48.2	45.0	46.9
	CyCTR [133] (NIPs 2021)	38.9	43.0	39.6	39.8	40.3	41.1	48.9	45.2	47.0	45.6
	ASGNet [62] (CVPR 2021)	-	-	-	-	34.6	-	-	-	-	42.5
	CMN [124] (ICCV 2021)	37.9	44.8	38.7	35.6	39.3	42.0	50.5	41.0	38.9	43.1
	CWT [78] (ICCV 2021)	32.2	36.0	31.6	31.6	32.9	40.1	43.8	39.0	42.4	41.3
	MFGN [125] (MDPI 2022)	40.8	45.5	41.1	39.1	41.6	46.1	52.3	46.2	44.3	47.2
	VAT [41] (ECCV 2022)	39.0	43.8	42.6	39.7	41.3	44.1	51.1	50.2	46.1	47.9
	BAM [58] (CVPR 2022)	43.4	50.6	47.5	43.4	46.2	49.3	54.2	51.6	49.6	51.2
	IPMT [73] (NeurIPS 2022)	41.4	45.1	45.6	40.0	43.0	43.5	49.7	48.7	47.9	47.5
	DCAMA [97] (ECCV 2022)	41.9	45.1	44.4	41.7	43.3	45.9	50.5	50.7	46.0	48.3
	FECANet [68] (TMM 2022)	38.5	44.6	42.6	40.7	41.6	44.6	51.5	48.4	45.8	47.6
	ABCNet [45] (CVPR 2023)	42.3	46.2	46.0	42.0	44.1	45.5	51.7	52.6	46.4	49.1
	PMNet [16] (WACV 2024)	39.8	41.0	40.1	40.7	40.4	50.1	50.0	49.6	49.6	50.3
ResNet101	FWB [84] (ICCV 2019)	17.0	18.0	21.0	28.9	21.2	19.1	21.5	23.9	30.1	23.7
	DAN [111] (ECCV 2020)	-	-	-	-	24.4	-	-	-	-	29.6
	PFENet [105] (TPAMI 2020)	36.8	41.8	38.7	36.7	38.5	40.4	46.8	43.2	40.5	42.7
	HSNet [81] (ICCV 2021)	37.2	44.1	42.4	41.3	41.2	45.9	53.0	51.8	47.1	49.5
	VPI [110] (WACV 2021)	24.9	25.8	21.9	21.1	23.4	27.8	29.3	24.7	29.4	27.8
	CWT [78] (ICCV 2021)	30.3	36.6	30.5	32.2	32.4	38.5	46.7	39.4	43.2	42.0
	API [103] (Pattern Recognition 2023)	33.9	46.6	24.8	31.2	31.2	35.7	47.5	31.6	35.9	35.9
	IPMT [73] (NeurIPS 2022)	40.5	45.7	44.8	39.3	42.6	45.1	50.3	49.3	46.8	47.9
Swin-B	DCAMA [97] (ECCV 2022)	41.5	46.2	45.2	41.3	43.5	48.0	58.0	54.3	47.1	51.9
	PMNet [16] (WACV 2024)	44.7	44.3	44.0	41.8	43.7	52.6	53.3	53.5	52.8	53.1
Swin-B	HSNet [81] (ICCV 2021)	43.6	49.9	49.4	46.4	47.3	50.1	58.6	56.7	55.1	55.1
	DCAMA [97] (ECCV 2022)	49.5	52.7	52.8	48.7	50.9	55.4	60.3	59.9	57.5	58.3

Table 4. Evaluation of 1-shot and 5-shot semantic segmentation specialist models on COCO-20ⁱ using FB-IoU scores available in the literature.

Backbone	Method	1 Shot	5 Shot
VGG16	PANet [113] (ICCV 2019)	59.2	63.5
	PFENet [105] (TPAMI 2020)	60	61.6
	SAGNN [121] (CVPR 2021)	61.2	63.1
	FECANet [68] (TMM 2022)	65.5	67.7
ResNet50	ASGNet [62] (CVPR 2021)	60.4	66.9
	HSNet [81] (ICCV 2021)	68.2	70.7
	VAT [41] (ECCV 2022)	68.8	72.4
	CMN [123] (ICCV 2021)	61.7	63.3
	MFGN [125] (MDPI 2022)	65.2	69.1
	CMN [123] (ICCV 2021)	61.7	63.3
	ASNet [50] (CVPR 2022)	68.8	71.6
	FECANet [68] (TMM 2022)	69.6	71.1
ResNet101	DAN [111] (ECCV 2020)	62.3	63.9
	PFENet [105] (TPAMI 2020)	63	65.8
	HSNet [81] (ICCV 2021)	69.1	72.4
	MFGN [125] (MDPI 2022)	65.6	69.1
	SAGNN [121] (CVPR 2021)	60.9	63.4
	A-MCG [43] (AAAI 2019)	52	54.7
Swin-B	HSNet [81] (ICCV 2021)	72.5	76.1
	DCAMA [97] (ECCV 2022)	73.2	76.9

Table 5. Quantitative comparison results of semantic segmentation specialist models on FSS-1000 dataset with the `mIoU` scores available in the literature.

Backbone	Method	1 Shot	5 Shot
VGG16	OSLSM [96] (BMVC 2017)	70.3	73
	co-FCN [89] (ICLR 2018)	71.2	74.2
	FSS-1000 [63] (CVPR 2020)	73.4	80.1
	DoG-LSTM [3] (WACV 2021)	80.8	83.4
	HSNet [81] (ICCV 2021)	82.3	85.8
	API [103] (arXiv 2021)	83.4	85.3
	DCAMA [97] (ECCV 2022)	88.2	88.8
ResNet50	FOMAML [39] (arXiv 2019)	75.1	80.6
	FSOT [71] (TMM 2022)	82.5	83.8
	HSNet [81] (ICCV 2021)	85.5	87.8
	VAT [41] (ECCV 2022)	89.5	90.3
	ASNet [50] (CVPR 2022)	68.8	71.6
	DCAMA [97] (ECCV 2022)	92.4	93.1
	PMNet [16] (WACV 2024)	84.6	86.3
ResNet101	DAN [111] (ECCV 2020)	85.2	88.1
	HSNet [81] (ICCV 2021)	86.5	88.5
	VAT [41] (ECCV 2022)	90	90.6
	API [103] (arXiv 2021)	85.6	88
	DCAMA [97] (ECCV 2022)	88.3	89.1
Swin-B	HSNet [81] (ICCV 2021)	86.7	88.9
	DCAMA [97] (ECCV 2022)	90.1	90.4

Table 6. Quantitative comparison between some modern specialist and generalist models on PASCAL-5ⁱ, COCO-20ⁱ, and FSS-1000. The results report the mean-IoU scores available in the literature.

Method	PASCAL-5 ⁱ		COCO-20 ⁱ		FSS-1000	
	one-shot	few-shot	one-shot	few-shot	one-shot	few-shot
<i>specialist model</i>						
HSNet [81] (ICCV 2021)	67.3	71.6	47.3	55.1	86.7	88.9
DCAMA [97] (ECCV 2022)	69.3	74.9	50.9	58.3	90.1	90.4
VAT [41] (ECCV 2022)	67.9	72.0	41.3	47.9	90.0	90.6
PMNet [16] (WACV 2024)	68.1	73.9	43.7	53.1	84.6	86.3
<i>generalist model</i>						
Painter [114] (CVPR 2023)	64.5	64.6	32.8	32.6	61.7	62.3
SegGPT [115] (ICCV 2023)	83.2	89.8	56.1	67.9	85.6	89.3

4.4 Qualitative samples

In the FSS field, researchers often present their models with various combinations of backbones and training sets, leading to diverse strengths and weaknesses for each combination. While quantitative results, as presented in the tables of Section 4.3, offer valuable insights into model performance, they may not fully convey the nuanced capabilities of different model configurations. To provide a more intuitive understanding of how these combinations influence a model’s effectiveness and to facilitate qualitative comparisons between different models, we present Table 7. This table showcases the one-shot predictions of selected models under various configurations while reporting the achieved %IoU score achieved by each prediction, contextualizing the quantitative results reported in Section 4.3.

From this qualitative analysis, notable observations emerge. For instance, it becomes evident that certain classes pose greater challenges for prediction, while some models demonstrate higher resilience to disparities between support and query images. A closer examination of specific rows in the table reveals insightful patterns. Notably, in the second row, all models struggle to differentiate between cars and buses. Furthermore, disparities in appearance from the support to the query image, such as those seen in the chair and person in rows 3 and 8, disproportionately affect the performance of certain models. Here is also evident how the generalist model SegGPT effectively leverages its larger and composite pretraining dataset, in many cases predicting higher quality masks with respect to its specialist counterparts, casting light on a possible direction for this research field.

In conclusion, qualitative analysis of predictions is indispensable for assessing the reliability of a model in specific applications or under particular challenges. Such analysis can unveil insights that may be challenging to glean solely from quantitative results.

Table 7. Qualitative 1-shot prediction samples from a selection of models. The first two columns show the support image with the support mask overlayed in red; the second column shows the query image along with the ground truth mask in overlay; the rest of the table shows the prediction by the selected models. For each prediction of the specialist models, the top row indicates the name of the model, the second reports the name of the backbone, and the third one reports the dataset used for training. For the generalist model SegGPT is not possible to talk about a backbone, and its training set is composed of a large combination of datasets. Different columns for a single model allow comparing predictions when using different combinations of backbones and training datasets. Under each prediction is reported the %IoU with respect of the ground truth.

Support	Query	DCAMA						HSNe		VAT			SegGPT						
		ResNet101		ResNet50		Swin-B		ResNet101		ResNet101		ResNet50							
		COCO	PASCAL	COCO	PASCAL	COCO	PASCAL	COCO	PASCAL	FSS-1000	PASCAL	COCO							
								8.25	0.76	17.73	11.10	8.14	8.00	40.73	79.90	65.83	13.34	54.31	85.50
								79.45	0.05	84.18	0.00	2.53	2.78	71.91	2.67	93.86	85.63	70.52	89.85
								31.40	21.20	25.69	33.23	20.39	7.31	2.33	37.96	16.86	37.63	37.27	80.51
								11.50	17.54	0.00	1.37	42.98	10.87	67.70	86.52	86.71	88.88	74.21	89.80
								25.59	15.12	17.78	0.00	17.12	17.07	38.16	54.00	48.09	48.97	49.98	90.07
								4.70	5.72	0.00	0.00	8.02	13.89	25.12	28.04	31.02	13.21	34.95	55.25
								0.00	0.09	14.65	0.32	14.29	0.53	56.00	68.81	72.91	69.63	56.04	79.87
								0.00	0.11	26.79	12.32	11.77	17.64	0.00	42.06	0.00	51.01	2.74	80.11

5 OUTLOOK IN FEW SHOT SEGMENTATION

In this section, we are going to highlight the limitations and criticisms of the FSS research line and models. We are also going to showcase proposed extensions of the presented FSS problem that could benefit other research fields.

Limitations and open challenges of episodic training

Episodic training, which is widely used and accepted as a good approach for FSS, has some drawbacks as noted by Cao et al. in [12]. Indeed, in real-world applications it can be difficult to predict and fix the number of available shots K as it ranges between applications. Cao et al. also theoretically demonstrated how this mismatch between the value of K during meta-training and meta-testing can lead to a decline in performance. Additionally, they showed that a model based on prototypes does not significantly improve its prediction accuracy by simply increasing the number of labelled samples in the support set.

Given these limitations, Boudiaf et al. [7] proposed the use of transductive learning. Unlike inductive learning, transductive learning techniques have access to all data, both training and testing, beforehand. The key idea of this approach is that during the learning process it is possible to make use of the patterns and additional information present in the testing data, even without knowing the labels. In such a way it could be possible to better exploit the few annotated examples typical of FSS.

N-Way K-Shot

A great difference between FSS and traditional supervised semantic segmentation is that FSS models learn to segment only one class, while most modern semantic segmentation models are capable of segmenting many different classes. To address this limitation, Dong et al. in [24] generalized the FSS problem to N classes naming it N-Way K-Shot Semantic Segmentation. In this setting we can define **N-Way K-Shot Semantic Segmentation (N-Way K-Shot SS)** as follow:

Definition 4 (N-Way K-Shot Semantic Segmentation). N-Way K-Shot Semantic Segmentation is the problem of predicting the semantic segmentation mask \hat{M}_q of the N different classes of the label-set L in a query image I_q given a support set S and K examples for each class.

More formally, defined a label set L of N classes:

$$L = \{l_i\}_{i=1}^N$$

we can build a support set:

$$S = \{(I^j, Y_{l_i}^j)\}_{j=1, i=1}^{j=K, i=N}$$

where is allowed $I^i = I^j$ as long as $Y^i \neq Y^j$. Is worth pointing out that in N-Way K-Shot FSS more than one class subject can be present in a given image. In addition, in this extension of FSS, also the output of the model is changed. Since the model is segmenting N classes plus the background, given a query image I_q with resolution $[H^q, W^q]$, the output of the model will have the shape of $[H^q, W^q, N + 1]$.

Generalized Few Shot Segmentation

Tian et al. in [104] highlight the limitations of traditional FSS models, which can only predict novel classes present in the support set and cannot predict base classes used for pretraining. They also stress that support and query images must contain the same classes for the model to perform optimally, meaning that if the query image does not contain the

subject of the support set, the model may not perform optimally and could potentially misbehave. In other words, the model expects to find the same objects in both the support set and query image, but this is not always guaranteed in real-life situations.

To address these limitations, Tian et al. in [104] and Lu et al. [77] envision a segmentation model that can both segment base classes learned from a sufficient number of samples and learn new classes in a few-shot manner. This problem is defined as **Generalized Few-shot Semantic Segmentation (GFS-Seg)** in their work. A potential difficulty of this setting is that the unbalanced number of training examples for the base and novel classes could lead to biases in the segmentation model. On this aspect, Liu et al. [70] speculate that as the base classes are learned with a higher number of examples, and are thus better learned, their prototypes should not be updated as much as the prototypes for novel classes which are learned from scratch. Therefore, they propose discouraging large modifications to base class prototypes while boosting large distances between prototypes with a class contrastive loss and a class relationship loss. With a similar intent, the Project onto Orthogonal Prototype (POP) [69] framework builds a set of orthogonal prototypes where the prototypes relative to the base classes remain frozen during meta training.

Aiming at a more real-world-oriented model, Hajimiri et al. [35] propose to extend the concept introduced by Tian et al. [104]. Starting from **GFS-Seg**, they incorporate the capability to predict a multi class segmentation mask, and not just a binary mask typical of many works in **FSS**. Their approach allows for the prediction of multiple classes on the query image, including both the base classes learned during pretraining and the novel classes learned in few-shot mode. Effectively bridging the gap between **GFS-Seg** and N-Way K-Shot Semantic Segmentation.

Lei et al. [60] identify yet another crucial gap in the deployment of Few-Shot Segmentation (FSS) models for real-world applications, notably the inherent assumption that base and novel classes originate from the same dataset domain, typically PASCAL-5ⁱ or COCO-20ⁱ.

However, this assumption fails to hold in practical scenarios, prompting the extension of **FSS** into **Cross-Domain Few-Shot Semantic Segmentation (CF-FSS)**. **CF-FSS** aims at crafting more versatile models capable of addressing diverse domain gaps. To this end, Lei et al. introduce a benchmark dataset comprising four distinct domains: FSS-100 [63] for everyday object images, Deepglobe [22] offering satellite images, ISIC2018 [20, 107] featuring dermoscopic images of skin lesions, and the Chest X-Ray dataset [11, 48] containing X-ray images.

This diverse dataset facilitates the implementation of **CF-FSS** training protocol, involving the division of the dataset into D_{train} and D_{test} to sample training episode and test episode, respectively. Crucially, D_{train} and D_{test} are designed to not have any overlap in terms of both image content and segmented classes, ensuring a substantial domain gap. The choice of subsets for D_{train} and D_{test} introduces varying degrees of domain shift, quantified by the authors with **Fréchet Inception Distance (FID)** [40].

This extension of FSS has spurred the development of robust models resilient to domain gaps, with subsequent works [16, 28, 46, 60] emphasizing the importance of translating domain-specific learned features into domain-agnostic representations while preserving the ability to discern target-specific features.

Incremental Few Shot Segmentation

With similar premises of **GFS-Seg**, Siam et al. in [98] and Carmelli et al. in [14] proposes to extend the **FSS** problem by setting it as incremental learning. Their rationale is that in applications such as robotics, once a segmentation model is deployed it might over time need to learn new classes while keeping the previously learned. Cermelli et al. named this problem **Incremental Few Shot Segmentation (iFSS)**. The main difference with **GFS-Seg** is that in iFSS the base dataset used to pretrain the model on base classes is no longer available to the model, which can access only the

data of the last episode to learn new classes and update the older. More recently Zhou et al. [139] highlight that in iFSS semantic confusion between base and novel classes might diminishes segmentation accuracy. They propose a method for semantic-guided relation alignment and adaptation to address this issue. Initially, they align base class representations with their semantic information, then conduct semantic-guided adaptation during incremental learning to ensure consistency between visual and semantic embeddings of encountered novel classes, thus reducing semantic overlap.

To simulate such a scenario Siam et al. used the PASCAL dataset in an incremental manner, naming it iPASCAL, to provide the model with two new classes per episode. The model they developed for this task is based on prototypical learning with a key difference: it keeps a rolling average for the known class prototypes and makes new class prototypes for the new classes. Using this approach their model is capable of keeping and updating the internal representation of the base classes while also being able of learning new ones.

6 CONCLUSIONS

Semantic Segmentation is a crucial computer vision task that has a wide range of applications, from healthcare to robotics and surveillance. However, a major challenge in applying state-of-the-art models to these fields is the requirement for a large labelled training dataset. Collecting and labelling such an amount of data is both expensive and time-consuming, and it can also raise privacy concerns as well as technical issues. Recently, the field of Few Shot Learning (FSL) has emerged to tackle the problem of learning a new task with limited training samples. This paper aims at examining the current challenges in applying FSL to the task of semantic segmentation, known as Few Shot Semantic Segmentation (FSS), and at providing a general overview of the main approaches found in the literature of this topic. Most works in this area can be categorised based on the main strategy they adopt to solve the task. We identified the major approaches to be Conditional Networks, Prototypical Networks, and Latent Space Optimization.

Conditional Networks were the first solution deployed for FSS, the challenge with these models is determining the best parameter set for the conditioning branch to pass to the segmentation branch. The parameter set must be informative enough to recognize the target class in the query image while allowing for appearance variation.

Prototypical Networks use a module to compute class representative prototypes, which serve as references to evaluate if a pixel in the query image belongs to a certain class. The main challenges with this approach are the metric used to compute the similarity between a pixel and the class prototype, as well as how to compute the prototype in order to make it informative enough for describing a whole class.

Latent Space Optimization approaches study how a well-structured latent space can aid in FSS. Some works train a GAN on the target class to produce a full-sized dataset, while others describe classes as statistical distributions in the latent space. The main limitation of this approach is that results are heavily dependent on the difficult training procedures of GANs.

Furthermore, we delineate how contemporary foundational models can be tailored for the Few Shot Semantic Segmentation (FSS), elucidating the potential of prompt engineering and multimodal models to pave the way for novel research trajectories. Additionally, we underscore the significance of generalist models, which possess the capacity to address diverse tasks, including the Few Shot Semantic Segmentation (FSS).

We also present a comprehensive overview of benchmark datasets, metrics, and significant results in the field of FSS summarizing the results of the most important works in this area, while also presenting a selection of qualitative results for the task of One Shot Semantic Segmentation.

Finally, this paper addresses some of the challenges and potential developments in FSS. Some researchers have argued that the traditional approach of episodic training may not be optimal, as mismatches between the number of shots available for meta-training and meta-testing can result in significant performance degradation. We explore N-Way K-Shot Semantic Segmentation as the task of learning to segment multiple classes with limited shots per class. Other authors have emphasized the importance of not forgetting the classes used during pretraining, and the need to continuously learn new classes, which has led to the development of **Generalized Few-shot Semantic Segmentation (GFS-Seg)** and **Incremental Few Shot Segmentation (iFSS)** methods.

In conclusion, the field of FSS is still open and active, and we can expect to see new and consequential works in the upcoming years. Ongoing developments and new approaches are expected to address the challenges in this area enabling the use of semantic segmentation models in new fields.

ACKNOWLEDGMENTS

This paper has been partially supported by “Agritech: Centro Nazionale per lo sviluppo delle nuove tecnologie in agricoltura” project funded by the European Union NextGenerationEU program within the PNRR.

REFERENCES

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*.
- [2] Wei Ao, Shunyi Zheng, Yan Meng, and Yang Yang. 2024. Few-shot semantic segmentation via mask aggregation. *Neural Processing Letters* (2024).
- [3] Reza Azad, Abdur R Fayjie, Claude Kauffmann, Ismail Ben Ayed, Marco Pedersoli, and Jose Dolz. 2021. On the texture bias for few-shot cnn segmentation. In *IEEE/CVF Winter conference on Applications of Computer Vision (WACV)*.
- [4] Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems* (2019).
- [5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* (2017).
- [6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
- [7] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. 2021. Few-shot segmentation without meta-learning: A good transductive inference is all you need?. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* (2020).
- [9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Conference on Neural Information Processing Systems (NeurIPS)*.
- [10] Attila Budai, Rüdiger Bock, Andreas Maier, Joachim Hornegger, Georg Michelson, et al. 2013. Robust vessel segmentation in fundus images. *International journal of biomedical imaging* (2013).
- [11] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P. Musco, Rahul Kumar Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Kiran Antani, George R. Thoma, and Clement J. McDonald. 2014. Lung Segmentation in Chest Radiographs Using Anatomical Atlases With Nonrigid Registration. *IEEE Transactions on Medical Imaging* (2014).
- [12] Tianshi Cao, Marc Law, and Sanja Fidler. 2019. A theoretical analysis of the number of shots in few-shot learning. *arXiv preprint arXiv:1909.11722* (2019).
- [13] Nico Catalano, Alessandro Maranelli, Agnese Chiatti, and Matteo Matteucci. 2024. More than the Sum of Its Parts: Ensembling Backbone Networks for Few-Shot Segmentation. *arXiv preprint arXiv:2402.06581* (2024).
- [14] Fabio Cermelli, Massimiliano Mancini, Yongqin Xian, Zeynep Akata, and Barbara Caputo. 2021. Prototype-based Incremental Few-Shot Semantic Segmentation. In *British Machine Vision Conference (BMVC)*.

- [15] Zhaobin Chang, Yonggang Lu, Xingcheng Ran, Xiong Gao, and Xiangwen Wang. 2023. Few-shot semantic segmentation: a review on recent approaches. *Neural Computing and Applications* (2023).
- [16] Hao Chen, Yonghan Dong, Zheming Lu, Yunlong Yu, and Jungong Han. 2024. Pixel Matching Network for Cross-Domain Few-Shot Segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
- [17] Jiacheng Chen, Bin-Bin Gao, Zongqing Lu, Jing-Hao Xue, Chengjie Wang, and Qingmin Liao. 2021. Apanet: adaptive prototypes alignment network for few-shot semantic segmentation. *arXiv preprint arXiv:2111.12263* (2021).
- [18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*.
- [19] Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [20] Noel C. F. Codella, Veronica M Rotemberg, Philipp Tschandl, M. E. Celebi, Stephen W. Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyritis, Michael Armando Marchetti, Harald Kittler, and Allan C. Halpern. 2019. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). *ArXiv* (2019).
- [21] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [22] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- [23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [24] Nanqing Dong and Eric P. Xing. 2018. Few-Shot Semantic Segmentation with Prototype Learning. In *British Machine Vision Conference (BMVC)*.
- [25] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. 2014. Discriminative Unsupervised Feature Learning with Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*.
- [26] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2012. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- [27] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. (2018).
- [28] Haoran Fan, Qi Fan, Maurice Pagnucco, and Yang Song. 2023. DARNet: Bridging Domain Gaps in Cross-Domain Few-Shot Segmentation with Dynamic Adaptation. *ArXiv abs/2312.04813* (2023).
- [29] Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. 2022. Self-support Few-Shot Semantic Segmentation. In *European Conference on Computer Vision (ECCV)*.
- [30] Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R. Rudnicka, Christopher G. Owen, and Sarah A. Barman. 2012. An Ensemble Classification-Based Approach Applied to Retinal Blood Vessel Segmentation. *IEEE Transactions on Biomedical Engineering* (2012).
- [31] Siddhartha Gairola, Mayur Hemani, Ayush Chopra, and Balaji Krishnamurthy. 2020. SimPropNet: Improved Similarity Propagation for Few-shot Image Segmentation. In *International Joint Conference on Artificial Intelligence (IJCAI)*.
- [32] Alberto Garcia-Garcia, Sergio Orts-Escalano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. 2018. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing* (2018).
- [33] Albert Gu and Tri Dao. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752* (2023).
- [34] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. 2018. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval* (2018).
- [35] Sina Hajimiri, Malik Boudiaf, Ismail Ben Ayed, and Jose Dolz. 2023. A Strong Baseline for Generalized Few-Shot Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [36] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. 2022. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336* (2022).
- [37] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. 2014. Simultaneous detection and segmentation. In *European Conference on Computer Vision (ECCV)*.
- [38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [39] Sean M Hendryx, Andrew B Leach, Paul D Hein, and Clayton T Morrison. 2019. Meta-learning initializations for image segmentation. *arXiv preprint arXiv:1912.06290* (2019).
- [40] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Neural Information Processing Systems*.
- [41] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. 2022. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision (ECCV)*.

- [42] A.D. Hoover, V. Kouznetsova, and M. Goldbaum. 2000. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging* (2000).
- [43] Tao Hu, Pengwan Yang, Chiliang Zhang, Gang Yu, Yadong Mu, and Cees GM Snoek. 2019. Attention-based multi-context guiding for few-shot semantic segmentation. In *AAAI conference on artificial intelligence*.
- [44] Gabriel Huang, Issam Laradji, David Vazquez, Simon Lacoste-Julien, and Pau Rodriguez. 2022. A survey of self-supervised and few-shot object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- [45] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. 2023. Rethinking Federated Learning With Domain Shift: A Prototype View. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [46] Xinyang Huang, Chuang Zhu, and Wenkai Chen. 2023. RestNet: Boosting Cross-Domain Few-Shot Segmentation with Residual Transformation Network. In *British Machine Vision Conference (BMVC)*.
- [47] Ehtesham Iqbal, Sirojbek Safarov, and Seongdeok Bang. 2022. MSANet: Multi-Similarity and Attention Guidance for Boosting Few-Shot Segmentation. *arXiv preprint arXiv:2206.09667* (2022).
- [48] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les R. Folio, Jenifer Siegelman, Fiona M. Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul Kumar Singh, Sameer Kiran Antani, George R. Thoma, Yixiang Wang, Pu-Xuan Lu, and Clement J. McDonald. 2014. Automatic Tuberculosis Screening Using Chest Radiographs. *IEEE Transactions on Medical Imaging* (2014).
- [49] Suvarna Kadam and Vinay Vaidya. 2020. Review and analysis of zero, one and few shot learning approaches. In *Intelligent Systems Design and Applications (ISDA)*.
- [50] Dahyun Kang and Minsu Cho. 2022. Integrative few-shot learning for classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [51] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [52] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [53] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [54] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- [55] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
- [56] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *arXiv:2304.02643* (2023).
- [57] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *ArXiv abs/1411.2539* (2014).
- [58] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. 2022. Learning what not to segment: A new perspective on few-shot segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [59] Chunbo Lang, Binfei Tu, Gong Cheng, and Junwei Han. 2022. Beyond the Prototype: Divide-and-conquer Proxies for Few-shot Segmentation. In *International Joint Conference on Artificial Intelligence (IJCAI)*.
- [60] Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. 2022. Cross-Domain Few-Shot Semantic Segmentation. In *European Conference on Computer Vision (ECCV)*.
- [61] Biao Li, Yong Shi, Zhiqian Qi, and Zhensong Chen. 2018. A Survey on Semantic Segmentation. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*.
- [62] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. 2021. Adaptive prototype learning and allocation for few-shot segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [63] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. 2020. Fss-1000: A 1000-class dataset for few-shot segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [64] Yiwen Li, Gratiavianus Wesley Putra Data, Yunguan Fu, Yipeng Hu, and Victor Adrian Prisacariu. 2021. Few-shot Semantic Segmentation with Self-supervision from Pseudo-classes. *British Machine Vision Conference (BMVC)* (2021).
- [65] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. 2019. Look into Person: Joint Body Parsing & Pose Estimation Network and a New Benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).
- [66] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*.
- [67] Binghao Liu, Jianbin Jiao, and Qixiang Ye. 2021. Harmonic Feature Activation for Few-Shot Semantic Segmentation. *IEEE Transactions on Image Processing* (2021).
- [68] Huafeng Liu, Pai Peng, Tao Chen, Qiong Wang, Yazhou Yao, and Xian-Sheng Hua. 2023. Fecanet: Boosting few-shot semantic segmentation with feature-enhanced context-aware network. *IEEE Transactions on Multimedia* (2023).
- [69] Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. 2023. Learning Orthogonal Prototypes for Generalized Few-Shot Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

- [70] Weide Liu, Zhonghua Wu, Yang Zhao, Yuming Fang, Chuan-Sheng Foo, Jun Cheng, and Guosheng Lin. 2023. Harmonizing Base and Novel Classes: A Class-Contrastive Approach for Generalized Few-Shot Segmentation. *ArXiv* (2023).
- [71] Weide Liu, Chi Zhang, Henghui Ding, Tzu-Yi Hung, and Guosheng Lin. 2022. Few-shot Segmentation with Optimal Transport Matching and Message Flow. *IEEE Transactions on Multimedia* (2022).
- [72] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. 2020. Crnet: Cross-reference networks for few-shot segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [73] YUANWEI LIU, Nian Liu, Xiwen Yao, and Junwei Han. 2022. Intermediate Prototype Mining Transformer for Few-Shot Semantic Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [74] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. 2020. Part-aware prototype network for few-shot semantic segmentation. In *European Conference on Computer Vision (ECCV)*.
- [75] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. 2024. Matcher: Segment Anything with One Shot Using All-Purpose Feature Matching. In *The Twelfth International Conference on Learning Representations (ICLR)*.
- [76] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [77] Zhihe Lu, Sen He, Da Li, Yi-Zhe Song, and Tao Xiang. 2023. Prediction Calibration for Generalized Few-Shot Semantic Segmentation. *IEEE Transactions on Image Processing* (2023).
- [78] Zhihe Lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. 2021. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [79] Timo Lüddecke and Alexander Ecker. 2022. Image Segmentation Using Text and Image Prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [80] Lingchen Meng, Shiyi Lan, Hengduo Li, Jose M. Alvarez, Zuxuan Wu, and Yu-Gang Jiang. 2023. SEGIC: Unleashing the Emergent Correspondence for In-Context Segmentation. *ArXiv* (2023).
- [81] Juhong Min, Dahyun Kang, and Minsu Cho. 2021. Hypercorrelation squeeze for few-shot segmentation. In *IEEE/CVF international conference on computer vision (ICCV)*.
- [82] Tom M Mitchell and Tom M Mitchell. 1997. *Machine learning*. McGraw-hill New York.
- [83] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. *Foundations of machine learning*. MIT press.
- [84] Khoi Nguyen and Sinisa Todorovic. 2019. Feature weighting and boosting for few-shot segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.
- [85] Atsuro Okazawa. 2022. Interclass Prototype Relation for Few-Shot Segmentation. In *European Conference on Computer Vision (ECCV)*.
- [86] Maxime Oquab, Timothée Darcret, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research* (2024).
- [87] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR.
- [88] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. (2018).
- [89] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alexei A Efros, and Sergey Levine. 2018. Few-shot segmentation propagation with guided networks. *arXiv preprint arXiv:1806.07373* (2018).
- [90] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. 2023. PACO: Parts and Attributes of Common Objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [91] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning Deep Representations of Fine-Grained Visual Descriptions. In *2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [92] Wenqi Ren, Yang Tang, Qiyu Sun, Chaoqiang Zhao, and Qing-Long Han. 2022. Visual Semantic Segmentation Based on Few/Zero-Shot Learning: An Overview. *arXiv preprint arXiv:2211.08352* (2022).
- [93] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML)*.
- [94] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI)*.
- [95] Oindrila Saha, Zezhou Cheng, and Subhransu Maji. 2022. GANORCON: Are Generative Models Useful for Few-shot Segmentation?. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [96] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. 2017. One-Shot Learning for Semantic Segmentation. In *British Machine Vision Conference (BMVC)*.
- [97] Xinyu Shi, Dong Wei, Yu Zhang, Donghuan Lu, Munan Ning, Jiashun Chen, Kai Ma, and Yefeng Zheng. 2022. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *European Conference on Computer Vision (ECCV)*.

- [98] Mennatullah Siam and Boris Oreshkin. 2019. Adaptive masked weight imprinting for few-shot segmentation. (2019).
- [99] Mennatullah Siam, Boris Oreshkin, and Martin Jagersand. 2019. AMP: Adaptive Masked Proxies for Few-Shot Segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.
- [100] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*.
- [101] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems* (2017).
- [102] J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, and B. van Ginneken. 2004. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging* (2004).
- [103] Haoliang Sun, Xiankai Lu, Haochen Wang, Yilong Yin, Xiantong Zhen, Cees G.M. Snoek, and Ling Shao. 2023. Attentional prototype inference for few-shot segmentation. *Pattern Recognition* (2023).
- [104] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. 2022. Generalized few-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [105] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. 2020. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence* (2020).
- [106] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. 2021. Repurposing gans for one-shot semantic part segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [107] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* (2018).
- [108] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* (2017).
- [109] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems*.
- [110] Haochen Wang, Yandan Yang, Xianbin Cao, Xiantong Zhen, Cees Snoek, and Ling Shao. 2021. Variational prototype inference for few-shot semantic segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
- [111] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. 2020. Few-shot semantic segmentation with democratic attention networks. In *European Conference on Computer Vision (ECCV)*.
- [112] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. 2021. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. In *Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS)*.
- [113] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. 2019. Panet: Few-shot image semantic segmentation with prototype alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [114] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. 2023. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [115] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. 2023. SegGPT: Towards Segmenting Everything In Context. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.
- [116] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (CSUR)* (2020).
- [117] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. 2022. CRIS: CLIP-Driven Referring Image Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [118] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. 2019. iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- [119] Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. 2021. Learning meta-class memory for few-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [120] Pengfei Xian, Lai-Man Po, Yuzhi Zhao, Wing-Yin Yu, and Kwok-Wai Cheung. 2023. CLIP Driven Few-Shot Panoptic Segmentation. *IEEE Access* (2023).
- [121] Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. 2021. Scale-aware graph neural network for few-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [122] Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao. 2021. Few-Shot Semantic Segmentation With Cyclic Memory Network. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.
- [123] Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao. 2021. Few-Shot Semantic Segmentation with Cyclic Memory Network. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.
- [124] Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao. 2021. Few-shot semantic segmentation with cyclic memory network. In *IEEE/CVF International Conference on Computer Vision (2021)*.
- [125] Chenjing Xin, Xinfu Li, and Yunfeng Yuan. 2022. Multilevel Features-Guided Network for Few-Shot Segmentation. *MDPI Electronics* (2022).

- [126] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. 2020. Prototype mixture models for few-shot semantic segmentation. In *European Conference on Computer Vision (ECCV)*.
- [127] Xianghui Yang, Bairun Wang, Kaise Chen, Xinchi Zhou, Shuai Yi, Wanli Ouyang, and Luping Zhou. 2020. Brinet: Towards bridging the intra-class and inter-class gaps in one-shot segmentation. *arXiv preprint arXiv:2008.06226* (2020).
- [128] Yuwei Yang, Fanman Meng, Hongliang Li, Qingbo Wu, Xiaolong Xu, and Shuai Chen. 2020. A new local transformation module for few-shot segmentation. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*.
- [129] Bingfeng Zhang, Jimin Xiao, and Terry Qin. 2021. Self-Guided and Cross-Guided Learning for Few-Shot Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [130] Bingfeng Zhang, Jimin Xiao, and Terry Qin. 2021. Self-guided and cross-guided learning for few-shot segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [131] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. 2019. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [132] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. 2019. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [133] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. 2021. Few-shot segmentation via cycle-consistent transformer. *Advances in Neural Information Processing Systems* (2021).
- [134] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. 2024. Personalize Segment Anything Model with One Shot. In *The Twelfth International Conference on Learning Representations (ICLR)*.
- [135] Xiaolin Zhang, Yunchao Wei, Zhao Li, Chenggang Yan, and Yi Yang. 2022. Rich Embedding Features for One-Shot Semantic Segmentation. *IEEE Transactions on Neural Networks and Learning Systems* (2022).
- [136] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. 2020. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE transactions on cybernetics* (2020).
- [137] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. 2021. Datasetgan: Efficient labeled data factory with minimal human effort. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [138] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [139] Yuan Zhou, Xin Chen, Yanrong Guo, Jun Yu, Richang Hong, and Qi Tian. 2024. Advancing Incremental Few-Shot Semantic Segmentation via Semantic-Guided Relation Alignment and Adaptation. In *MultiMedia Modeling*.
- [140] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. 2023. ZegCLIP: Towards adapting CLIP for zero-shot semantic segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023).
- [141] Kai Zhu, Wei Zhai, and Yang Cao. 2020. Self-Supervised Tuning for Few-Shot Segmentation. In *International Joint Conference on Artificial Intelligence (IJCAI)*.
- [142] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. *arXiv preprint arXiv:2401.09417* (2024).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009