Memory-augmented Variational Adaptation for Online Few-shot Segmentation

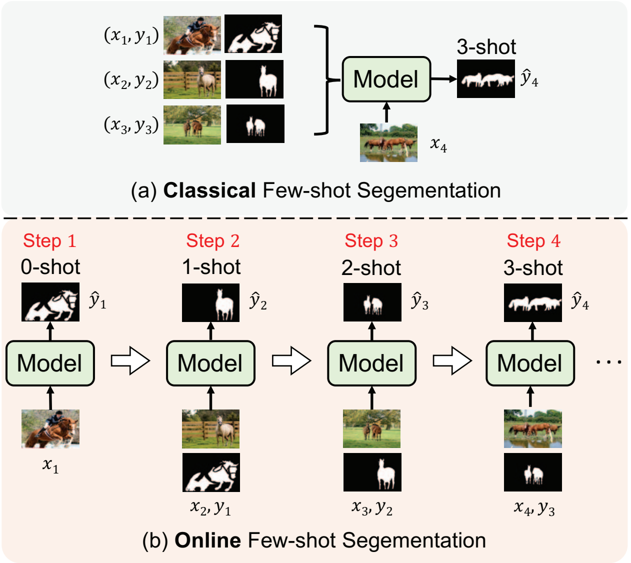
Jie Liu[1], Yingjun Du[1], Zehao Xiao[1], Cees G.M Snoek[1], Jan-Jakob Sonke[2], Efstratios Gavves[1]
[1]University of Amsterdam, Netherlands [2]The Netherlands Cancer Institute, Netherlands
[1]{j.liu5, y.du, z.xiao, cgmsnoek, egavves}@uva.nl [2]j.sonke@nki.nl

Abstract

In this paper, we investigate online few-shot segmentation, which learns to make mask prediction for novel classes while observing samples sequentially. The main challenge in such an online scenario is the sample diversity in the sequence, resulting in models learned from previous samples that do not generalize well to future samples. To this end, we propose a memory-augmented variational adaptation network, which learns to adapt the model to each new sample that arrives sequentially. Specifically, we first introduce a contextual prototypical memory, which retains category knowledge from previous contextual information to facilitate the model adaptation to future samples. The adaptation to each new sample is then formulated as a variational Bayesian inference problem, which strives to generate sample-specific model parameters by conditioning the sample and the prototypical memory. Furthermore, we propose a feature customization module to learn sample-specific feature representation for better model adaptation to each sample in the sequence. With extensive experiments, we show that the proposed method effectively adapts to each sample from the online sample sequence, thus achieving state-of-the-art performance on both natural image and medical image datasets.

1. Introduction

Recent advances in few-shot semantic segmentation (FSS) [25, 5, 36, 31, 33, 19, 28, 21, 38, 17] has achieved great progress for the semantic segmentation task in data-scarcity scenarios. Generally, classical FSS (Figure 1 (a)) learns to segment objects from previously unseen classes, by providing models with a small set of annotated examples simultaneously, i.e., the *support set*. Yet, acquiring and storing multiple annotated samples simultaneously in the dynamic world is an unrealistic requirement, especially when the number of novel classes and annotated samples increases over time. Incremental FSS [2] attempts to tackle class-incremental few-shot segmentation task, i.e., the number of novel classes increase over time. Compared with class-incremental learning, sample-incremental learning is perhaps even more common in the real world, i.e., samples of novel classes usually emerge in an online manner. For instance, physicians collect new tumor images from a patient during the treatment process [18, 20], while simultaneously tumor segmentation mask for the new image usually becomes available after the last treatment step. An ideal FSS model is required to learn from online sample streams and adapt to the segmentation of new samples dynamically.



(a) **Classical** Few-shot Segementation

(b) **Online** Few-shot Segementation

Figure 1. **Comparison between (a) classical few-shot segmentation and (b) online few-shot segmentation under 1-way 3-shot setting.** (a) Given three annotated support images simultaneously, classical few-shot segmentation learns to predict mask for the query image. However, samples and annotation from novel classes usually emerge sequentially in our dynamic world, acquiring multiple annotated samples simultaneously is unrealistic. (b) online few-shot segmentation learns to make mask prediction for samples arriving sequentially, while corresponding mask annotation arrives afterwards.

In this work, we investigate the online few-shot segmentation task [1], which aims to make pixel-wise prediction for novel classes with samples arriving sequentially. We clarify the task with an example in Figure 1 (b), where the model is asked to segment the *goat* in the sequence. Specifically, the model can only access one sample and make prediction for it at a time step, while the corresponding ground truth is revealed to the model at the next time step. In such a way, the model evaluation and updating proceed alternately, and the model learns to segment samples from novel classes in an online manner. Ideally, the performance of online few-shot segmentation models should increase with the number of samples increases and exhibit smaller performance fluctuations. However, the main challenge in such an online scenario [8, 1] is the sample diversity in the sequence. Samples in one sequence comes from the same class but exhibit large appearance and scale variations. This results in models learned from previous samples that do not generalize well to future samples. In such case, achieving effective model adaptation to each sample in the sequence is essential for the online few-shot segmentation task.

To improve model adaptation capacity, we propose a memory-augmented variational adaptation network (MaVAN) for the online few-shot segmentation task. MaVAN is composed of three key components, i.e., contextual prototype memory, variational adaptation, and a feature customization module. First, we propose in Section 3.3 a contextual prototypical memory module to exploit task-specific contextual information . The memory retains category knowledge from previous samples and serves as dynamic support set for the segmentation of future samples. New class prototypes are approximated using groups of same-class exemplar embedding in the current sequence and stored contextual information in the external memory. Second, we propose in section 3.4 variational adaptation using a latent variable model in which we treat the classifier as a latent variable. We incorporate the category knowledge from the contextual prototypical memory and sample-specific context from the current sample to generate a probabilistic sample-specific classifier. We formulate the optimization of our variational adaptation as a variational inference problem by deriving a new evidence lower bound (ELBO) under the online setting. In doing so, the probabilistic classifier obtained are more informative and therefore better represent categories of objects compared to the deterministic vector. Third, we propose in Section 3.5 feature customization module to learn sample-specific feature representations better adapted to each sample. By doing so, the model is endowed with the ability to provide sample-specific segmentation for each sample in the sequence and copes with sample diversity well. Once trained on seen classes, our model could adapt to each sample from novel classes in the sequence with just a feed-forward computation at test time.

To sum up, our main contributions are as follows:

• We propose a memory-augmented variational adaptation network (MaVAN) to improve model adaption capacity for online few-shot segmentation. The contextual prototype memory is proposed to work as dynamic support set and retain task-specific context information for future samples.

• We formulate the model adaptation for each sample as a variational inference problem to generates sample-specific classifiers by deriving a new ELBO under the online setting.

• We propose a feature customization module to learn sample-specific feature representation for better model adaptation in the feature space.

• The proposed MaVAN achieves state-of-the-art performance for the online few-shot segmentation task on both natural image and medical image datasets.

2. Related work

Few-shot segmentation Given few support images with pixel-wise annotation, few-shot segmentation (FSS) aims to make mask prediction for the query image from novel classes. Existing FSS methods can be roughly divided into prototype-based methods [25, 5, 39, 36, 31, 33, 19, 28] and graph-based methods [30, 35, 21, 38]. Prototype-based methods usually generate prototypes from the support set with mask average pooling [25, 31, 36, 28] or K-means clustering [19], then these prototypes are used to interact with query feature to fuse cross-image context information. Graph-based methods adopt dense matching between masked support images and query image to excavate intrinsic similarity of different instances from the same class. For instance, HSNet [21] leverages 4D convolutions to achieve dense comparison reasoning over multi-level feature correlation between the support and query features. Recently, some work tend to extent classical FSS to more realistic settings, e.g., generalized FSS (GFSS) [27], incremental FSS (iFSS) [2], and online FSS [1] (OFSS). GFSS models learn to segment both old and new classes, while iFSS learns to segment both old and new classes with few samples without access to past training data. OFSS is first proposed in [1] to segment novel classes within online data streams with distractors. In this paper, we focus on the OFSS setting without distractors, i.e., intersection filed between few-shot segmentation and online learning .

Online learning Learning from a sequence of data instances dynamically, online learning aims to maximize the correctness for the sequence of predictions [11]. Various approaches, such as linear models [3], non-linear models with kernels [12, 15], and deep neural networks [40], have been proposed to tackle the online learning task. Recently, some online meta-learning methods [8, 1, 22] are proposed to tackle the online few-shot learning task, which aims to recognize novel classes from a sequence of data. The main challenge in this task is how to achieve faster and more efficient model adaptation to the new data by leveraging previously seen data. Finn et al. [8] propose to achieve fast model adaption to new data with a data buffer storing all task data. Babu et al. [1] design a layer-distributed memory network to learn fast adaption. Inspired by the above methods, we focus on the online few-shot segmentation task and propose an adaptation network.

Model adaptation Model adaptation plays an important role in some computer vision tasks, e.g., active learning [23, 34], continual learning [37, 4, 6], domain adaptation [16, 32]. [37] achieves model adaptation to novel classes in continual segmentation by proposing a representation compensation module, which decouples the representation learning of both old and new knowledge. [1] adopts prototype-based knowledge distillation to enforce model adaptation to novel classes without forgetting old classes. [37] proposes a test-time adaptation approach which adapts off-the-shelf source pretrained models to continually changing target data. For the online few-shot segmentation task, we claim that model adaptation is also crucial because of large sample diversity in the sequential data.

3. Methodology

3.1. Problem Statement

Classical few-shot semantic segmentation: Classical few-shot semantic segmentation follows the meta-learning paradigm, where a task (or episode) is composed of a support set S and a query set Q. Here, we consider the 1-way k-shot setting. Conditioned on the support set with k annotated support samples, the few-shot learner $f(\cdot)$ is expected to make pixel-wise prediction for the query sample x^q: $\hat{y}^q = f(x^q; (x_1^s, y_1^s), \ldots, (x_k^s, y_k^s))$, where x is input image, y is corresponding binary mask. However, this setup is built on the assumption that annotated support examples are revealed to the model simultaneously, which is usually unrealistic in our dynamic world, especially in medical.

Online few-shot semantic segmentation. Online few-shot semantic segmentation aims to make pixel-wise prediction on a stream of samples from novel classes. A task consists of T samples from the same novel class. In this setup, samples are revealed to the model sequentially, while corresponding masks are given afterwards. The few-shot learner in online few-shot segmentation aims to tackle the sequential decision problem: $\hat{y}_t = f(x_t; (x_1, \text{null}), (x_2, y_1), \ldots, (x_t, y_{t-1}))$, where *null* indicates no mask for the first sample, and the model makes a random prediction for x_1. Normally, we set $t > 1$ for illustration in the following text. By feeding sequential samples and subsequent labels to the model, we evaluate the model online while updating model parameters dynamically.

3.2. Model

We propose a memory-augmented variational adaptation network (MaVAN) for online few-shot segmentation. The proposed MaVAN achieves online few-shot segmentation via three key components: 1) A **Contextual prototypical memory** that retains category knowledge from previous samples to facilitate model adaptation to future samples. 2) **Variational test-time adaptation** which formulates the model adaptation to new samples as a variational Bayesian inference problem. 3) **Feature customization module** that learns sample-specific feature representation for better model adaptation to each sample in the feature space. The pipeline of our model is shown in Figure 2.

3.3. Contextual prototypical memory

Online few-shot segmentation involves segmenting samples from the same classes sequentially. Therefore, effectively leveraging acquired category knowledge from previous samples to boost the segmentation of future samples is crucial. Here, we construct a contextual prototypical memory to achieve this goal. Considering memory efficiency, we choose to represent sample prototypes p rather than original samples in the memory. Specifically, a sample $x \in \mathbb{R}^{3\times H\times W}$ with ground-truth $y \in \mathbb{R}^{H\times W}$ can be represented by a sample prototype $p \in \mathbb{R}^{N\times C}$, which is composed of N prototypes with C channels, respectively. Specifically, given a deep neural network $\Phi : \mathcal{X} \rightarrow \mathcal{Z}$, which maps from input space to feature space, the sample prototype p corresponds to the clustering centers of foreground features:

$$p = \mathcal{A}(\Phi(x) \odot y) = \mathcal{A}(z \odot y), \tag{1}$$

where $\mathcal{A}$ is a clustering function (e.g., K-means), and $z \in \mathbb{R}^{C\times H\times W}$ is the sample feature of H and W height and the width, respectively. We adopt element-wise multiplication between z and y to generate foreground features. At the time step t, the ground-truth y_{t-1} of sample x_{t-1} is revealed to the model so that we can store the sample prototype p_{t-1} of sample x_{t-1} into the memory. Similarly, we can store prototypes of all previous samples in the memory $\mathcal{M}_t = \{p_i\}_{i=1}^{t-1}$ sequentially. The contextual prototypical memory, which stores prototypes of previous samples sequentially, works as dynamic support set for the segmentation of future samples. Given the contextual prototypical memory aggregating category knowledge (e.g., different appearance of objects) from previous samples, we continue with a variational test-time adaptation in section 3.4 to achieve model adaption to future samples.

3.4. Variational test-time adaptation

Sample diversity in the sequence, e.g., large object appearances, is the main challenge for online few-shot segmentation, resulting in models learned from previous samples do not generalize well to future samples. Given appearance can change significantly over time in an online setting, we adopt a variation Bayesian model of classifier weights and integrate over all possible appearances, instead of making a point estimation for a classifier that learns to

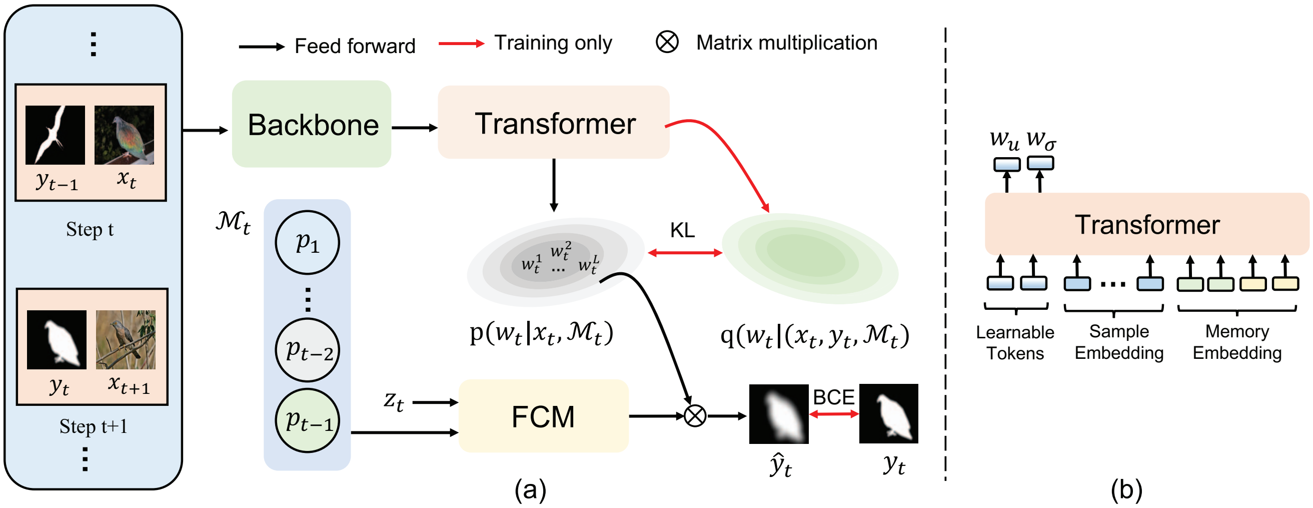


Figure 2. **a) Overview of the proposed memory-augmented variational adaptation network (MaVAN).** At t-th time step, the sample x_t and the previous sample's label y_{t-1} are revealed to the model, which first stores sample prototype p_{t-1} into the contextual prototypical memory $\mathcal{M}_t$ (section 3.3). Then, the model generates distributions of classifiers via transformer to achieve variational adaption (section 3.4) to the current sample . Lastly, sample feature z_t is feed into the feature customization module (FCM, section 3.5) to generate sample-specific feature $\hat{z}_t$, which further multiplies sample classifiers $\{w_t^1, w_t^2, ..., w_t^L\}$ from prior distribution $p(w_t|x_t, \mathcal{M}_t)$ to get predicted masks $\hat{y}_t$. **b) Details of distribution generation with transformer.** Learnable tokens interact with sample and memory embedding in transformer to generate distribution parameterized by w_μ and w_σ.

recognize only what it has seen the latest. Our online few-shot segmentation model is composed of a frozen backbone, a decoder network, and a classifier w. At time step t, rather than generating all model parameters, which is computation-expensive, we generate sample-specific classifier weights w_t for the sample x_t, maximizing the conditional predictive log-likelihood $\log p(y_t|x_t, \mathcal{M}_t)$. By incorporating the sample-specific classifier w_t into the predictive distribution, we have

$$\begin{aligned}&\log p(y_t|x_t, \mathcal{M}_t)\\&= \log \int p(y_t|x_t, w_t)p(w_t|\mathcal{M}_t)dw_t,\end{aligned} \tag{2}$$

where $p(w_t|\mathcal{M}_t)$ denotes the conditional prior distribution over w_t. By depending on the contextual prototypical memory $\mathcal{M}_t$, we infer the classifier w_t aggregating category knowledge from previous samples.

Although the contextual prototypical memory provides some category information, the model still knows little about the current sample, especially when previous and current samples exhibit large appearance variations. In other words, while the prior distribution $p(w_t|\mathcal{M}_t)$ preserves information from past samples, it might not be optimized for the current sample x_t we wish to analyze. To this end, we incorporate specific information about the current sample x_t, we further include x_t to the prior in Eq. (2), that is

$$\begin{aligned}&\log p(y_t|x_t, \mathcal{M}_t)\\&= \log \int p(y_t|x_t, w_t)p(w_t|x_t, \mathcal{M}_t)p(x_t)dw_t.\end{aligned} \tag{3}$$

Conditioned on the current sample x_t and the external memory $\mathcal{M}_t$, the prior distribution $p(w_t|x_t, \mathcal{M}_t)$ aggregates category knowledge from external memory and sample-specific knowledge from the current sample. To guarantee that the prior distribution $p(w_t|x_t, \mathcal{M}_t)$ could generate sample-specific classifier parameters, we design a variational posterior distribution $q(w_t|x_t, y_t, \mathcal{M}_t)$. By incorporating $q(w_t|x_t, y_t, \mathcal{M}_t)$ into Eq. (3), we derive a lower bound of the conditional predictive log-likelihood:

$$\begin{aligned}&\log p(y_t|x_t, \mathcal{M}_t)\\&= \log \int p(y_t|x_t, w_t)p(w_t|x_t, \mathcal{M}_t)p(x_t)dw_t\\&\geq \mathbb{E}_{q(w_t|x_t, y_t, \mathcal{M}_t)}[\log p(y_t|x_t, w_t)]\\&- \mathbb{D}_{\mathrm{KL}}[q(w_t|x_t, y_t, \mathcal{M}_t)||p(w_t|x_t, \mathcal{M}_t)].\end{aligned} \tag{4}$$

This formulation establishes a variational lower bound for the predictive distribution, which we can optimize. That way, we guarantee the inferred classifiers to be discriminative and adaptive to the segmentation of the current sample. Besides, the KL divergence term in Eq. (4) further works as a regularizer, pushing the prior distribution to adapt better to the current sample. In practice, we generate the prior and the posterior distribution via a vanilla transformer [29]

$$\begin{aligned}[w_{\mu_t}, w_{\sigma_t}] = \texttt{Transformer}(&\texttt{cls}_{\mu_t}, \texttt{cls}_{\sigma_t}, [x_t, x_{t-1}, \cdots, x_1],\\&[p_{t-1}, p_{t-2}, \cdots, p_0])\end{aligned} \tag{5}$$

to allow the flexibility of variable input sizes of conditions, where $[\texttt{cls}_{\mu_t}, \texttt{cls}_{\sigma_t}]$ are the mean and variance of classifier token embedding, respectively. The derivation of Eq. (4) is provided in the supplementary material. In addition to a sample-specific classifier, we also need a better feature representation (section 3.5) to achieve sample-specific segmentation.

3.5. Feature customization module

With variation adaptation, we generate sample-specific classifier for each sample, yet semantic segmentation requires much contextual information in the feature space to make a precise pixel-wise prediction. In such case, we propose a feature customization module, which strives to learn better representations with the contextual prototypical memory and the current sample. At time step t, we have a sample x_t, initial feature representation z_t, and the contextual prototypical memory $\mathcal{M}_t$. Specifically, we incorporate the object context from the contextual prototypical memory by introducing a category prototype p_t^c:

$$p_t^c = \frac{1}{t-1}\sum_{i=1}^{t-1} m_i p_i, \tag{6}$$

where

$$m_i = 1 - \frac{-y_i \log \hat{y}_i}{\sum\limits_{j=1}^{t-1} -y_k \log \hat{y}_j}. \tag{7}$$

m_i is the weight for the prototype p_i, derived from the prediction masks $\hat{y}_{t-1}$ at time step $t-1$, and the ground-truth masks y_{t-1} of the sample x_{t-1} revealed to the model at time t. A larger m_i implies higher confidence for model about the current segmentation. In turn, this implies that p_i has more relevant category knowledge in memory. The category prototypes p_t^c are updated per time step and expected to obtain robust and generalizable class representation with time goes on.

Segmentation in the online setting requires contextual information from both previous and current samples. TO the end, we obtain the updated feature representation $\hat{z}_t$ of sample x_t via a decoder network:

$$\hat{z}_t = \Psi([z_t, p_t^c, y_t^*]), \tag{8}$$

where Ψ is the decoder network implemented with multiple convolutional layers, p_t^c is the expanding variant of p_t^c with same spatial dimension as z_t. y_t^* is the pseudo mask of sample x_t modelled by pixel-wise cosine similarity between class prototype p_t^c and initial feature representation z_t, and $[\cdot]$ indicates the concatenation operation in the channel dimension. The initial feature representation z_t and prior mask y_t^* provide sample-specific context from current sample x_t, while the category prototype $\tilde{p}_t^c$ contains category knowledge from previous samples. In such a way, we learn sample-specific feature representation for better adaptation to the segmentation of each sample.

With sample-specific representation $\hat{z}_t$, we can directly make mask prediction for sample x_t with classifiers sampled from the prior distribution: $\hat{y}_t = \frac{1}{L}\sum_{l=1}^{L} \hat{z}_t w_t^l$, where $w_t^l \sim p(w_t|z_t, \mathcal{M}_t)$. L is number of Monte Carlo samples.

3.6. Meta-training and meta-test

In the meta-training stage, we sample sequences from base classes for model training. The loss function is computed after all the segmentation tasks in the sequence are completed. By incorporating feature representation z_t and $\hat{z}_t$ into the evidence lower bound in Eq. (4), the final objective function is formulated as:

$$\begin{aligned}\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}\Big[\frac{1}{L}\sum_{l=1}^{L}[-\log p(y_t|\hat{z}_t, w_t^l)] \\ + \mathbb{D}_{\mathrm{KL}}(q(w_t|z_t, y_t, \mathcal{M}_t))||p(w_t|z_t, \mathcal{M}_t))\Big],\end{aligned} \tag{9}$$

where T is the length of sequences. To enable back propagation, we adopt the reparameterization trick [14] for sampling the classifier w_t. In practice, the first log-likelihood term is implemented as a cross entropy loss between predictions and ground-truth. The conditional probabilistic distributions are set to be diagonal Gaussian. We implement them using multi-layer perceptrons (MLP) with the amortization technique and the reparameterization trick [14], which take the conditionals as input and output the parameters of the Gaussian.

In the meta-test stage, we sample sequences from novel classes for evaluation. Specifically, at a time step t, the model receives the current sample x_t and ground-truth y_{t-1} of previous sample as input from a test task. Then we directly sample classifiers from the prior distribution $p(w_t|x_t, \mathcal{M}_t)$ to make mask prediction $\hat{y}_t$ for the current sample x_t. Note that there is no backpropagation to update the model parameters, and only the contextual prototypical memory and the prior distribution will be updated.

4. Experiments

4.1. Experimental Setup

Datasets We adopt two natural image datasets, i.e., PASCAL [1] and COCO [2], and one medical dataset ABD-MRI-20 [3], to evaluate the performance of proposed method. PASCAL is created from PASCAL VOC 2012 [7] and additional SBD annotations [9]. It contains 20 classes, split into 15 training and 5 testing classes. COCO is a more challenging dataset, which is composed of 60 training classes and 20 testing classes. ABD-MRI-20 [13] is an MRI dataset, which contains 20 3D T2-SPIR MRI scans and each with

[1] http://host.robots.ox.ac.uk/pascal/VOC/

[2] https://cocodataset.org/#download

[3] https://chaos.grand-challenge.org/Data/

Settings	1-shot	2-shot	3-shot	4-shot	5-shot	mean
Deterministic classifier	54.25 ±1.75	55.93 ±1.65	58.23 ±2.38	56.96 ±2.30	59.41 ±1.17	56.95 ±0.23
Variational classifier	**54.84** ±**1.37**	**56.77** ±**1.73**	**58.96** ±**2.87**	**57.41** ±**2.37**	**60.03** ±**0.78**	**57.60** ±**0.18**

Table 1. **Variational vs. deterministic classifier in (%) on PASCAL with ResNet50 averaged three runs**. Variational classifier is more critical than the deterministic classifier.

Settings	1-shot	2-shot	3-shot	4-shot	5-shot	mean
Prototype-augmented	56.17 ±1.32	59.42 ±2.74	57.06 ±2.86	58.62 ±0.68	59.05 ±1.47	58.06 ±0.38
Memory-augmented	**59.41** ±**2.50**	**60.21** ±**1.67**	**62.83** ±**2.31**	**61.42** ±**0.73**	**62.90** ±**1.29**	**61.35** ±**0.14**

Table 2. **Benefit of memory-augmented segmentation in (%) on PASCAL with ResNet50 averaged three runs.** Memory-augmented segmentation achieves better performance than with prototype-augmented segmentation.

four organs, i.e., liver, spleen, and left and right kidney. We adopt 5 scans with the spleen for evaluation and the remaining 15 scans with other organs for training. T image-mask pairs are randomly sampled from a specific class to construct a data sequence. More details about datasets setup are provided in supplemental material.

Evaluation Metrics We adopt mIoU (mean Intersection over union) as a metric for evaluation on two natural image datasets and dice score on the medical dataset. Given an input sequence with length T, the model makes a random guess on the first image and outputs predicted masks for the remaining images in sequence. We compute the t-shot mIoU (or dice score), i.e., the performance after seeing t image-mask pairs in the sequence. We set the length of the input sequence T as 6 for both training and evaluation, and report 1-shot to 5-shot results . To characterize the learning ability of O-FSS models over sequences, we also present the averaged mIoU from 1-shot to 5-shot results. All numbers are reported with 1000 sequences for natural image datasets and 100 sequences for a medical dataset.

4.2. Baseline Models

We compare the model adaptation ability of the proposed model with three baseline models, i.e., online prototypical network (OPN) [22], LSTM [24], incremental few-shot segmentation model PIFS [2]. More details about the extension of above methods to the OFSS setting and comparisons with classical few-shot segmentation models care illustrated in the supplementary materials.

OPN Ren et al. [22] extend the Prototypical Network [26] to the online setting, where prototypes are updated sequentially using weighted averaging. To achieve model adaptation in the OFSS task, we adopt OPN to aggregate prototypes in the prototype memory into the category prototype.

LSTM [10] We include temporal modelling methods for comparison as well. Santoro et al. [24] utilize LSTM for the online few-shot learning task. Similarly, we adopt a single-layer LSTM to interact with the prototype memory and update the category prototype iteratively.

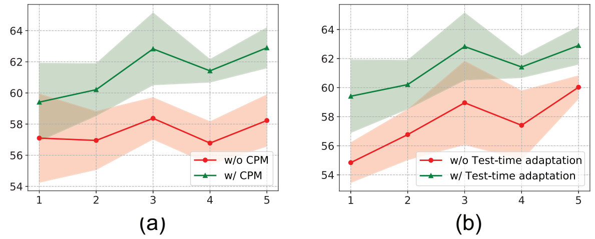


Figure 3. **Benefits of (a) contextual prototypical memory (CPM) and (b) test-time adaptation in (%) on PASCAL with ResNet50 averaged three runs.** Our model with prototype memory and test-time adaptation achieves consistently better performance across 1-shot to 5-shot than its correspondence variants.

PIFS Cermelli et al. [2] propose a prototype-based model adaptation network (PIFS) for incremental few-shot segmentation. We extend PIFS to the online few-shot segmentation task by introducing prototype-based distillation loss on both old and new sample prototypes.

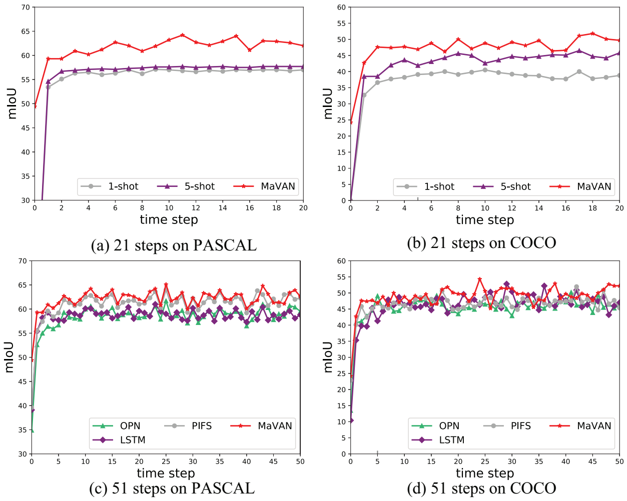
4.3. Results

Benefit of prototype memory To show the importance of the prototype memory, we implement a model variant without prototype memory. We directly replace the prototype memory in Eq. 4 with the sample prototype from last time step; that is, we utilize p_{t-1} to generate prior distribution $p(w_t|x_t, p_{t-1})$ and perform the segmentation task of the sample x_t. The experimental results are reported in Figure 3 (a). Without the prototype memory, our model has difficulty in aggregating category information from previous samples and adapting to new samples. Thus the performance of our model without prototype memory in all shots is worse than that in the memory-based model.

Importance of test-time adaptation We investigate the benefit of the test-time adaptation on the PASCAL dataset in Figure 3 (b). In this paper, the **memory adaptation** in Eq. 2 is without test-time adaptation, which is only conditioning on the prototype memory, while the **sample adaptation** in

Dataset	Method	1-shot	2-shot	3-shot	4-shot	5-shot	mean
PASCAL	OPN [22]	52.63 ± 3.74	55.87 ± 5.62	57.97 ± 2.61	56.47 ± 3.45	59.30 ± 1.89	56.45 ±0.20
	LSTM [24]	55.40 ± 2.79	57.37 ± 4.57	58.97 ± 3.36	57.37 ± 2.92	59.63 ± 0.74	57.74 ± 0.25
	PIFS [2]	57.09 ±2.03	**61.60 ±3.37**	58.83 ±2.16	60.25 ±1.43	60.66 ±1.52	59.69 ±0.41
	MaVAN	**59.41 ±2.50**	60.21 ±1.67	**62.82 ±2.31**	**61.42 ±0.73**	**62.90 ±1.29**	**61.35 ±0.23**
COCO	OPN [22]	39.59 ±1.87	44.37 ±2.26	42.60 ±1.40	42.53 ±2.62	45.22 ±0.36	42.86 ±1.27
	LSTM [24]	35.52 ±1.89	41.19 ±4.30	41.45 ±7.32	44.10 ±0.19	44.65 ±1.36	41.38 ±1.45
	PIFS [2]	40.15 ±1.13	45.83 ±2.72	42.45 ±2.53	45.12 ±2.36	46.73 ±1.53	44.06 ±1.38
	MaVAN	**43.08 ±1.61**	**47.57 ±2.28**	**45.96 ±1.18**	**46.71 ±5.83**	**49.17 ±3.24**	**46.50 ±1.16**
ABD-MRI	OPN [22]	35.40 ±1.27	39.72 ±0.33	30.95 ±0.42	34.73 ±0.11	36.86 ±0.27	35.53 ±0.76
	LSTM [24]	34.66 ±1.40	37.80 ±0.14	29.08 ±0.20	32.23 ±1.35	35.82 ±0.78	33.92 ±0.65
	PIFS [2]	38.19 ±0.63	42.32 ±0.33	31.78 ±0.20	36.49 ±0.51	38.07 ±0.85	37.37 ±0.69
	MaVAN	**39.57 ±0.58**	**44.94 ±0.46**	**34.48 ±0.12**	**38.90 ±0.67**	**41.26 ±1.88**	**39.83 ±0.83**

Table 3. **Comparison with baseline models on three datasets. PASCAL and COCO adopt mIoU as metric, while ABD-MRI-20 uses dice score, mean value and variance are reported with three runs.** Our model is a consistent top-performer on both natural image and medical image datasets, outperforming baseline models by a large margin.



(a) 21 steps on PASCAL

(b) 21 steps on COCO

(c) 51 steps on PASCAL

(d) 51 steps on COCO

Figure 4. Segmentation performance over long sequences. We compare with classical FSS baselines (trained under 1-shot and 5-shot settings) and online FSS baselines in the first and second row, respectively. For sequences with length 21 and 51, the proposed MaVAN consistently outperforms classical and online FSS methods on both PASCAL and COCO.

Eq. 3 is with test-time adaptation, which conditions on both the current sample and the prototype memory. From Figure 3 (b), we see that incorporating the variational classifier with test-time adaptation performs consistently better than that without test-time adaptation. This is because, with the test-time adaptation mechanism, our model can learn the capability to adapt to the segmentation of the current sample using sample-specific knowledge from current samples and category information from previous samples.

Variational vs. deterministic classifier We compare against the deterministic classifier as our baseline model in which few-shot segmentation training methods obtain the classifier. As shown in Table 1, the proposed variational classifier consistently outperforms the deterministic classifier, demonstrating the benefit brought by probabilistic modeling. The variational classifier provides more informative representations of classes, which are able to encompass large intra-class variations and, therefore, improve performance with time step increases.

Benefit of memory-augmented segmentation We demonstrate the benefits of memory-augmented segmentation on the PASCAL dataset. We implement a prototype-augmented variant of our model by replacing the category prototype in Eq. 6 with the sample prototype from the last time step. As shown in Table 2, our model with memory-augmented segmentation performs consistently better than that with prototype-augmented segmentation. The comparison clarifies that introducing category knowledge from prototype memory to representation learning is beneficial for better adaptation to the segmentation task of new samples.

Segmentation of long sequences We investigate model performance on long sequences by increasing time steps to 21 and 51, respectively. In Figure 4 (a) and (b), we compare with classical few-shot segmentation models (more details can be found in supplemental material) trained under 1-shot and 5-shot settings. Interestingly, a simple extension of classical few-shot segmentation does not cope well with sequential data loading, and tends to converge to over-smoothed, averaged masks of lesser accuracy. In Figure 4

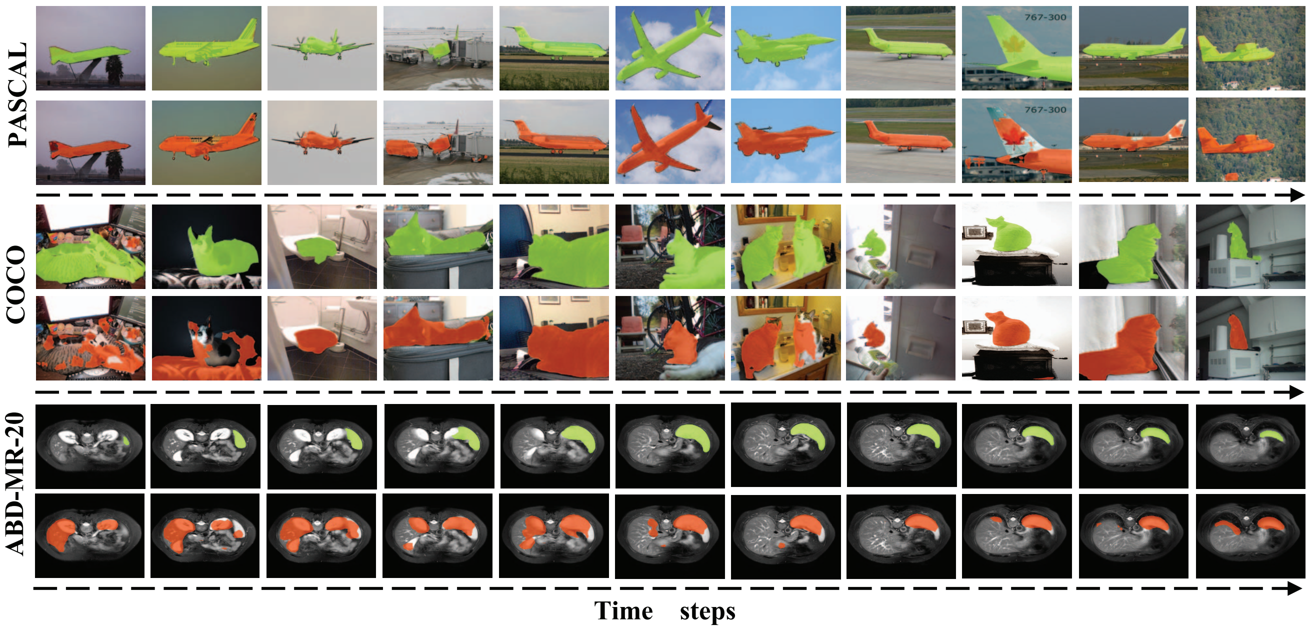


Figure 5. **Visualization of online few-shot segmentation performance on PASCAL (top), COCO (middle), and ABD-MRI-20 (bottom)**. Ground-truths are masked in green, while predictions are masked in red. Generally, our model could make better and better mask prediction as time steps increase, even though sometimes experiences fluctuation.

(c) and (d), we increase the time step to 50 and compare it with four online few-shot segmentation baseline models. Our model achieves superior performance than the four baseline variants with time step increases. Our model consistently outperforms classical few-shot segmentation models. This is because of the variational test-time adaptation mechanism, which dynamically adapts the model to new samples in the sequence.

Comparison with baseline models As shown in Table 3, the proposed method sets consistent state-of-the-art performance on all online few-shot segmentation benchmarks. On the PASCAL dataset, our model surpasses the second-best method, i.e., PIFS, by a margin of 1.66% in terms of mean mIoU. In addition, our model is also a consistent top-performer on the COCO dataset, outperforming other baseline methods by 2.44%$\sim$3.64% in terms of mean mIoU. This is reasonable since we generate model parameters with sample-specific knowledge from the current sample and category knowledge from previous samples, leading to more adapted models. Further, in the online medical segmentation task the proposed model still achieves best performance. We conclude that our model robustly improves online few-shot segmentation performance in both natural and medical scenarios. More detailed comparision results can be found in the supplementary materials.

Qualitative Results. In Figure 5, we report qualitative results from our model on both natural image and medical image datasets. At the first time step, our model gives a random guess on the target mask, as no auxiliary information about target is provided. With time step increases and mask annotation of previous is released, our model improves the segmentation performance of target object iteratively, even though exhibits some fluctuation. We can conclude that the proposed MaVAN achieves effective model adaptation to new samples, thus making more and more accurate mask prediction. More experimental results can be found in the supplemental material.

5. Conclusion

In this paper, we investigate online few-shot segmentation, which aims to make pixel-wise prediction for samples from novel classes sequentially. To cope with large sample diversity in the sequence, we propose a memory-augmented variational adaptation network MaVAN, which adapts model to each new sample. We first propose a contextual prototypical memory to retain category knowledge from previous samples, then formulate the model adaptation to the sample as a variational Bayesian inference problem. Conditioned on the current sample and an external memory, our method is able to generate sample-specific classifiers for the sample at each time step. Furthermore, we propose feature customization module to learn sample-specific representation for each sample. By doing so, our method is updated sequentially and achieves fast adaptation to each sample segmentation task with the number of samples increases over time. Extensive experiments on both natural image and medical datasets show

References

[1] Sudarshan Babu, Pedro Savarese, and Michael Maire. Online meta-learning via learning with layer-distributed memory. *Advances in Neural Information Processing Systems*, 34:14795–14808, 2021. 1, 2, 3

[2] Fabio Cermelli, Massimiliano Mancini, Yongqin Xian, Zeynep Akata, and Barbara Caputo. Prototype-based incremental few-shot semantic segmentation. *arXiv preprint arXiv:2012.01415*, 2020. 1, 2, 6, 7

[3] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006. 2

[4] Sungmin Cha, YoungJoon Yoo, Taesup Moon, et al. Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. *Advances in neural information processing systems*, 34:10919–10930, 2021. 3

[5] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, 2018. 1, 2

[6] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4040–4050, 2021. 3

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 5

[8] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *International Conference on Machine Learning*, pages 1920–1930. PMLR, 2019. 2

[9] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011. 5

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 6

[11] Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021. 2

[12] Rong Jin, Steven CH Hoi, and Tianbao Yang. Online multiple kernel learning: Algorithms and mistake bounds. In *International conference on algorithmic learning theory*, pages 390–404. Springer, 2010. 2

[13] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 5

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5

[15] Jyrki Kivinen, Alexander J Smola, and Robert C Williamson. Online learning with kernels. *IEEE transactions on signal processing*, 52(8):2165–2176, 2004. 2

[16] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9641–9650, 2020. 3

[17] Jie Liu, Yanqi Bao, Guo-Sen Xie, Huan Xiong, Jan-Jakob Sonke, and Efstratios Gavves. Dynamic prototype convolution network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2022. 1

[18] Qin Liu, Zhenlin Xu, Yining Jiao, and Marc Niethammer. isegformer: Interactive segmentation via transformers with application to 3d knee mr images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 464–474. Springer, 2022. 1

[19] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 142–158. Springer, 2020. 1, 2

[20] Xiangde Luo, Guotai Wang, Tao Song, Jingyang Zhang, Michael Aertsen, Jan Deprest, Sebastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning. *Medical Image Analysis*, 72:102102, 2021. 1

[21] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6941–6952, 2021. 1, 2

[22] Mengye Ren, Michael L Iuzzolino, Michael C Mozer, and Richard S Zemel. Wandering within a world: Online contextualized few-shot learning. *arXiv preprint arXiv:2007.04546*, 2020. 2, 6, 7

[23] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021. 3

[24] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016. 6, 7

[25] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017. 1, 2

[26] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 6

[27] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11563–11572, 2022. 2

[28] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1, 2

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[30] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *European Conference on Computer Vision*, pages 730–746. Springer, 2020. 2

[31] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9197–9206, 2019. 1, 2

[32] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 3

[33] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 763–778. Springer, 2020. 1, 2

[34] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 93–102, 2019. 3

[35] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9587–9595, 2019. 2

[36] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5217–5226, 2019. 1, 2

[37] Chang-Bin Zhang, Jia-Wen Xiao, Xialei Liu, Ying-Cong Chen, and Ming-Ming Cheng. Representation compensation networks for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7053–7064, 2022. 3

[38] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. Few-shot segmentation via cycle-consistent transformer. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2

[39] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics*, 50(9):3855–3865, 2020. 2

[40] Guanyu Zhou, Kihyuk Sohn, and Honglak Lee. Online incremental feature learning with denoising autoencoders. In *Artificial intelligence and statistics*, pages 1453–1461. PMLR, 2012. 2