Integrative Few-Shot Learning for Classification and Segmentation

Dahyun Kang   Minsu Cho
Pohang University of Science and Technology (POSTECH), South Korea
http://cvlab.postech.ac.kr/research/iFSL

Abstract

We introduce the integrative task of few-shot classification and segmentation (FS-CS) that aims to both classify and segment target objects in a query image when the target classes are given with a few examples. This task combines two conventional few-shot learning problems, few-shot classification and segmentation. FS-CS generalizes them to more realistic episodes with arbitrary image pairs, where each target class may or may not be present in the query. To address the task, we propose the integrative few-shot learning (iFSL) framework for FS-CS, which trains a learner to construct class-wise foreground maps for multi-label classification and pixel-wise segmentation. We also develop an effective iFSL model, attentive squeeze network (ASNet), that leverages deep semantic correlation and global self-attention to produce reliable foreground maps. In experiments, the proposed method shows promising performance on the FS-CS task and also achieves the state of the art on standard few-shot segmentation benchmarks.

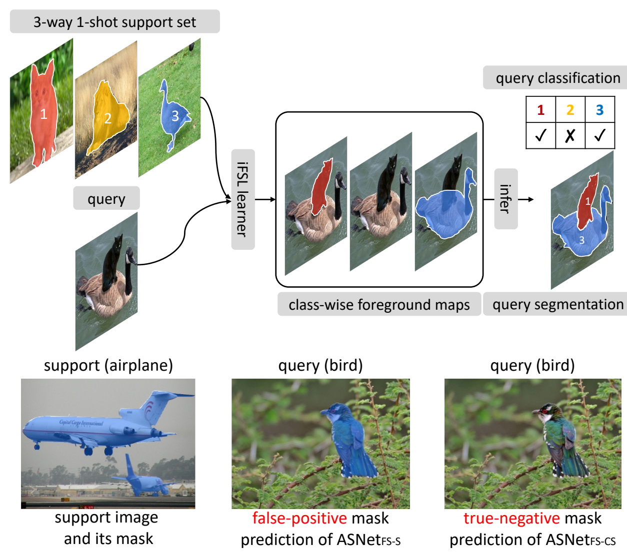

Figure 1. Top: Integrative few-shot learning framework (iFSL) for integrative few-shot classification and segmentation (FS-CS). **Bottom:** FS-S learners are trained to segment a query image using a semantically-coupled support set thus often blindly highlight any salient objects regardless of support semantics. The proposed FS-CS learners are trained to predict class presence as well as corresponding masks thus correctly discriminate what to segment based on the semantic relevance between the query and the support.

1. Introduction

Few-shot learning [15,16,30,82,84] is the learning problem where a learner experiences only a limited number of examples as supervision. In computer vision, it has been most actively studied for the tasks of image classification [22,29,66] and semantic segmentation [6,41,48,59] among many others [21,49,54,94,99]. Few-shot classification (FS-C) aims to classify a query image into target classes when a few support examples are given for each target class. Few-shot segmentation (FS-S) is to segment out the target class regions on the query image in a similar setup. While being closely related to each other [33,92,100], these two few-shot learning problems have so far been treated individually. Furthermore, the conventional setups for the few-shot problems, FS-C and FS-S, are limited and do not reflect realistic scenarios; FS-C [28,55,76] presumes that the query always contains one of the target classes in classification, while FS-S [25,52,63] allows the presence of multiple classes but does not handle the absence of the target classes in segmentation. These respective limitations prevent few-shot learning from generalizing to and evaluating on more realistic cases in the wild. For example, when a query image without any target class is given as in Fig. 1, FS-S learners typically segment out arbitrary salient objects in the query.

To address the aforementioned issues, we introduce the *integrative task of few-shot classification and segmentation* (FS-CS) that combines the two few-shot learning problems into a multi-label and background-aware prediction problem. Given a query image and a few-shot support set for target classes, FS-CS aims to *identify the presence of each target class* and *predict its foreground mask* from the query. Unlike FS-C and FS-S, it does not presume either the class exclusiveness in classification or the presence of all the target classes in segmentation.

As a learning framework for FS-CS, we propose *integrative few-shot learning* (iFSL) that learns to construct shared foreground maps for both classification and segmentation. It naturally combines multi-label classification and pixel-wise segmentation by sharing class-wise foreground maps and also allows to learn with class tags or segmentation annotations. For effective iFSL, we design the *attentive squeeze network* (ASNet) that computes semantic correlation tensors between the query and the support image features and then transforms the tensor into a foreground map by strided self-attention. It generates reliable foreground maps for iFSL by leveraging multi-layer neural features [44, 45] and global self-attention [11, 75]. In experiments, we demonstrate the efficacy of the iFSL framework on FS-CS and compare ASNet with recent methods [44, 86–88]. Our method significantly improves over the other methods on FS-CS in terms of classification and segmentation accuracy and also outperforms the recent FS-S methods on the conventional FS-S. We also cross-validate the task transferability between the FS-C, FS-S, and FS-CS learners, and show the FS-CS learners effectively generalize when transferred to the FS-C and FS-S tasks.

Our contribution is summarized as follows:

- We introduce the task of *integrative few-shot classification and segmentation* (FS-CS), which combines few-shot classification and few-shot segmentation into an integrative task by addressing their limitations.
- We propose the *integrative few-shot learning framework* (iFSL), which learns to both classify and segment a query image using class-wise foreground maps.
- We design the *attentive squeeze network* (ASNet), which squeezes semantic correlations into a foreground map for iFSL via strided global self-attention.
- We show in extensive experiments that the framework, iFSL, and the architecture, ASNet, are both effective, achieving a significant gain on FS-S as well as FS-CS.

2. Related work

Few-shot classification (FS-C). Recent FS-C methods typically learn neural networks that maximize positive class similarity and suppress the rest to predict the most probable class. Such a similarity function is obtained by a) meta-learning embedding functions [1,23,26,28,51,67,76,93,96], b) meta-learning to optimize classifier weights [17, 61, 69], or c) transfer learning [7,9,20,37,50,58,71,81], all of which aim to generalize to unseen classes. This conventional formulation is applicable if a query image corresponds to no less or more than a single class among target classes. To generalize FS-C to classify images associated with either none or multiple classes, we employ the multi-label classification [4, 8, 12, 31, 42]. While the conventional FS-C methods make use of the class uniqueness property via using the categorical cross-entropy, we instead devise a learning framework that compares the binary relationship between the query and each support image individually and estimates a binary presence of the corresponding class.

Few-shot semantic segmentation (FS-S). A prevalent FS-S approach is learning to match a query feature map with a set of support feature embeddings that are obtained by collapsing spatial dimensions at the cost of spatial structures [10, 18, 38, 40, 47, 65, 78, 89, 90, 95, 98]. Recent methods [73, 86–88, 97] focus on learning structural details by leveraging dense feature correlation tensors between the query and each support. HSNet [44] learns to squeeze a dense feature correlation tensor and transform it to a segmentation mask via high-dimensional convolutions that analyze the local correlation patterns on the correlation pyramid. We inherit the idea of learning to squeeze correlations and improve it by analyzing the spatial context of the correlation with effective global self-attention [75]. Note that several methods [68, 77, 91] adopt non-local self-attention [80] of the query-key-value interaction for FS-S, but they are distinct from ours in the sense that they learn to transform image feature maps, whereas our method focuses on transforming dense correlation maps via self-attention.

FS-S has been predominantly investigated as an one-way segmentation task, *i.e.*, foreground or background segmentation, since the task is defined so that every target (support) class object appears in query images, thus being not straightforward to extend to a multi-class problem in the wild. Consequently, most work on FS-S except for a few [10, 40, 70, 78] focuses on the one-way segmentation, where the work of [10,70] among the few presents two-way segmentation results from person-and-object images only, *e.g.*, images containing (person, dog) or (person, table).

Comparison with other few-shot approaches. Here we contrast FS-CS with other loosely-related work for generalized few-shot learning. Few-shot open-set classification [39] brings the idea of the open-set problem [14, 62] to few-shot classification by allowing a query to have no target classes. This formulation enables background-aware classification as in FS-CS, whereas multi-label classification is not considered. The work of [19, 72] generalizes few-shot segmentation to a multi-class task, but it is mainly studied under the umbrella of incremental learning [5, 43, 56]. The work of [64] investigates weakly-supervised few-shot segmentation using image-level vision and language supervision, while FS-CS uses visual supervision only. The aforementioned tasks generalize few-shot learning but differ from FS-CS in the sense that FS-CS integrates two related problems under more general and relaxed constraints.

3. Problem formulation

Given a query image and a few support images for target classes, we aim to *identify the presence* of each class and *predict its foreground mask* from the query (Fig. 1), which we call the *integrative few-shot classification and segmentation* (FS-CS). Specifically, let us assume a target (support) class set $\mathcal{C}_\text{s}$ of N classes and its support set $\mathcal{S} = \{(\mathbf{x}_\text{s}^{(i)}, y_\text{s}^{(i)}) | y_\text{s}^{(i)} \in \mathcal{C}_\text{s}\}_{i=1}^{NK}$, which contains K labeled instances for each of the N classes, *i.e.*, N-way K-shot [55, 76]. The label $y_\text{s}^{(i)}$ is either a class tag (weak label) or a segmentation annotation (strong label). For a given query image $\mathbf{x}$, we aim to identify the multi-hot class occurrence $\mathbf{y}_\text{C}$ and also predict the segmentation mask $\mathbf{Y}_\text{S}$ corresponding to the classes. We assume the class set of the query $\mathcal{C}$ is a subset of the target class set, *i.e.*, $\mathcal{C} \subseteq \mathcal{C}_\text{s}$, thus it is also possible to obtain $\mathbf{y}_\text{C} = \varnothing$ and $\mathbf{Y}_\text{S} = \varnothing$. This naturally generalizes the existing few-shot classification [67, 76] and few-shot segmentation [52, 63].

Multi-label background-aware prediction. The conventional formulation of few-shot classification (FS-C) [17, 67, 76] assigns the query to one class among the target classes exclusively and ignores the possibility of the query belonging to none or multiple target classes. FS-CS tackle this limitation and generalizes FS-C to multi-label classification with a background class. A multi-label few-shot classification learner f_C compares semantic similarities between the query and the support images and estimates class-wise occurrences: $\hat{\mathbf{y}}_\text{C} = f_\text{C}(\mathbf{x}, \mathcal{S}; \theta)$ where $\hat{\mathbf{y}}_\text{C}$ is an N-dimensional multi-hot vector each entry of which indicates the occurrence of the corresponding target class. Note that the query is classified into a *background* class if none of the target classes were detected. Thanks to the relaxed constraint on the query, *i.e.*, the query not always belonging to exactly one class, FS-CS is more general than FS-C.

Integration of classification and segmentation. FS-CS integrates multi-label few-shot classification with semantic segmentation by adopting pixel-level spatial reasoning. While the conventional FS-S [47, 52, 63, 65, 78] assumes the query class set exactly matches the support class set, *i.e.*, $\mathcal{C} = \mathcal{C}_\text{s}$, FS-CS relaxes the assumption such that the query class set can be a subset of the support class set, *i.e.*, $\mathcal{C} \subseteq \mathcal{C}_\text{s}$. In this generalized segmentation setup along with classification, an integrative FS-CS learner f estimates both class-wise occurrences and their semantic segmentation maps: $\{\hat{\mathbf{y}}_\text{C}, \hat{\mathbf{Y}}_\text{S}\} = f(\mathbf{x}, \mathcal{S}; \theta)$. This combined and generalized formulation gives a high degree of freedom to both of the few-shot learning tasks, which has been missing in the literature; the integrative few-shot learner can predict multi-label background-aware class occurrences and segmentation maps simultaneously under a relaxed constraint on the few-shot episodes.

4. Integrative Few-Shot Learning (iFSL)

To solve the FS-CS problem, we propose an effective learning framework, *integrative few-shot learning (iFSL)*. The iFSL framework is designed to jointly solve few-shot classification and few-shot segmentation using either a class tag or a segmentation supervision. The integrative few-shot learner f takes as input the query image $\mathbf{x}$ and the support set $\mathcal{S}$ and then produces as output the class-wise foreground maps. The set of class-wise foreground maps $\mathcal{Y}$ is comprised of $\mathbf{Y}^{(n)} \in \mathbb{R}^{H \times W}$ for N classes:

$$\mathcal{Y} = f(\mathbf{x}, \mathcal{S}; \theta) = \{\mathbf{Y}^{(n)}\}_{n=1}^{N}, \tag{1}$$

where $H \times W$ denotes the size of each map and θ is parameters to be meta-learned. The output at each position on the map represents the probability of the position being on a foreground region of the corresponding class.

Inference. iFSL infers both class-wise occurrences and segmentation masks on top of the set of foreground maps $\mathcal{Y}$. For class-wise occurrences, a multi-hot vector $\hat{\mathbf{y}}_\text{C} \in \mathbb{R}^N$ is predicted via max pooling followed by thresholding:

$$\hat{\mathbf{y}}_\text{C}^{(n)} = \begin{cases} 1 & \text{if } \max_{\mathbf{p} \in [H] \times [W]} \mathbf{Y}^{(n)}(\mathbf{p}) \geq \delta, \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

where $\mathbf{p}$ denotes a 2D position, δ is a threshold, and $[k]$ denotes a set of integers from 1 to k, *i.e.*, $[k] = \{1, 2, \cdots, k\}$. We find that inference with average pooling is prone to miss small objects in multi-label classification and thus choose to use max pooling. The detected class at any position on the spatial map signifies the presence of the class.

For segmentation, a segmentation probability tensor $\mathbf{Y}_\text{S} \in \mathbb{R}^{H \times W \times (N+1)}$ is derived from the class-wise foreground maps. As the background class is not given as a separate support, we estimate the background map in the context of the given supports; we combine N class-wise background maps into *an episodic background map* on the fly. Specifically, we compute the episodic background map $\mathbf{Y}_\text{bg}$ by averaging the probability maps of not being foreground and then concatenate it with the class-wise foreground maps to obtain a segmentation probability tensor $\mathbf{Y}_\text{S}$:

$$\mathbf{Y}_\text{bg} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{1} - \mathbf{Y}^{(n)}), \tag{3}$$

$$\mathbf{Y}_\text{S} = [\mathbf{Y} || \mathbf{Y}_\text{bg}] \in \mathbb{R}^{H \times W \times (N+1)}. \tag{4}$$

The final segmentation mask $\hat{\mathbf{Y}}_\text{S} \in \mathbb{R}^{H \times W}$ is obtained by computing the most probable class label for each position:

$$\hat{\mathbf{Y}}_\text{S} = \underset{n \in [N+1]}{\arg\max}\, \mathbf{Y}_\text{S}. \tag{5}$$

Learning objective. The iFSL framework allows a learner to be trained using a class tag or a segmentation annotation

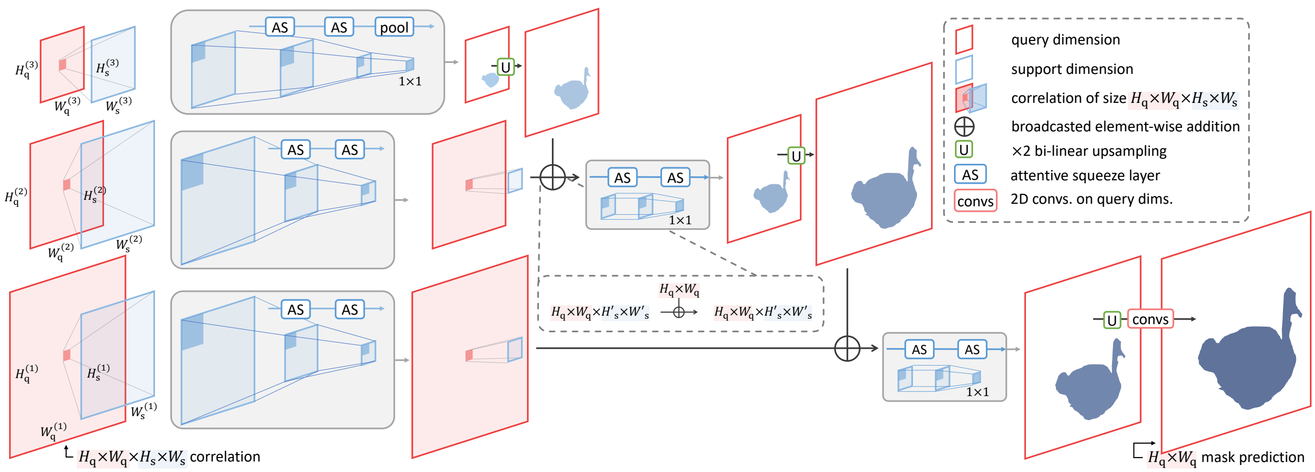

Figure 2. Overview of ASNet. ASNet first constructs a hypercorrelation [44] with image feature maps between a query (colored red) and a support (colored blue), where the 4D correlation is depicted as two 2D squares for demonstrational simplicity. ASNet then learns to transform the correlation to a foreground map by gradually squeezing the support dimension on each query dimension via global self-attention. Each input correlation, intermediate feature, and output foreground map has a channel dimension but is omitted in the illustration.

using the classification loss or segmentation loss, respectively. The classification loss is formulated as the average binary cross-entropy between the spatially average-pooled class scores and its ground-truth class label:

$$\mathcal{L}_{\text{C}} = -\frac{1}{N}\sum_{n=1}^{N} \mathbf{y}_{\text{gt}}^{(n)} \log \frac{1}{HW} \sum_{\mathbf{p}\in[H]\times[W]} \mathbf{Y}^{(n)}(\mathbf{p}), \tag{6}$$

where $\mathbf{y}_{\text{gt}}$ denotes the multi-hot encoded ground-truth class.

The segmentation loss is formulated as the average cross-entropy between the class distribution at each individual position and its ground-truth segmentation annotation:

$$\mathcal{L}_{\text{S}} = -\frac{1}{(N+1)}\frac{1}{HW}\sum_{n=1}^{N+1} \sum_{\mathbf{p}\in[H]\times[W]} \mathbf{Y}_{\text{gt}}^{(n)}(\mathbf{p}) \log \mathbf{Y}_{\text{S}}^{(n)}(\mathbf{p}), \tag{7}$$

where $\mathbf{Y}_{\text{gt}}$ denotes the ground-truth segmentation mask.

These two losses share a similar goal of classification but differ in whether to classify each *image* or each *pixel*. Either of them is thus chosen according to the given level of supervision for training.

5. Model architecture

In this section, we present *Attentive Squeeze Network* (ASNet) of an effective iFSL model. The main building block of ASNet is the attentive squeeze layer (AS layer), which is a high-order self-attention layer that takes a correlation tensor and returns another level of correlational representation. ASNet takes as input the pyramidal cross-correlation tensors between a query and a support image feature pyramids, *i.e.*, a hypercorrelation [44]. The pyramidal correlations are fed to pyramidal AS layers that gradually squeeze the spatial dimensions of the support image, and the pyramidal outputs are merged to a final foreground map in a bottom-up pathway [34,35,44]. Figure 2 illustrates the overall process of ASNet. The N-way output maps are computed in parallel and collected to prepare the class-wise foreground maps in Eq. (1) for iFSL.

5.1. Attentive Squeeze Network (ASNet)

Hypercorrelation construction. Our method first constructs NK hypercorrelations [44] between a query and each NK support image and then learns to generate a foreground segmentation mask w.r.t. each support input. To prepare the input hypercorrelations, an episode, *i.e.*, a query and a support set, is enumerated into a paired list of the query, a support image, and a support label: $\{(\mathbf{x}, (\mathbf{x}_{\text{s}}^{(i)}, y_{\text{s}}^{(i)}))\}_{i=1}^{NK}$. The input image is fed to stacked convolutional layers in a CNN and its mid- to high-level output feature maps are collected to build a feature pyramid $\{\mathbf{F}^{(l)}\}_{l=1}^{L}$, where l denotes the index of a unit layer, *e.g.*, `Bottleneck` layer in ResNet50 [22]. We then compute cosine similarity between each pair of feature maps from the pair of query and support feature pyramids to obtain 4D correlation tensors of size $H_{\text{q}}^{(l)} \times W_{\text{q}}^{(l)} \times H_{\text{s}}^{(l)} \times W_{\text{s}}^{(l)}$, which is followed by ReLU [46]:

$$\mathbf{C}^{(l)}(\mathbf{p}_{\text{q}}, \mathbf{p}_{\text{s}}) = \text{ReLU}\left(\frac{\mathbf{F}_{\text{q}}^{(l)}(\mathbf{p}_{\text{q}}) \cdot \mathbf{F}_{\text{s}}^{(l)}(\mathbf{p}_{\text{s}})}{||\mathbf{F}_{\text{q}}^{(l)}(\mathbf{p}_{\text{q}})|| \, ||\mathbf{F}_{\text{s}}^{(l)}(\mathbf{p}_{\text{s}})||}\right). \tag{8}$$

These L correlation tensors are grouped by P groups of the identical spatial sizes, and then the tensors in each group are concatenated along a new channel dimension to build a hypercorrelation pyramid: $\{\mathbf{C}^{(p)} | \mathbf{C}^{(p)} \in \mathbb{R}^{H_{\text{q}}^{(p)} \times W_{\text{q}}^{(p)} \times H_{\text{s}}^{(p)} \times W_{\text{s}}^{(p)} \times C_{\text{in}}^{(p)}}\}_{p=1}^{P}$ such that the channel size $C_{\text{in}}^{(p)}$ corresponds to the number of concatenated tensors in the p_{th} group. We denote the first two spatial dimensions of the correlation tensor, *i.e.*, $\mathbb{R}^{H_{\text{q}} \times W_{\text{q}}}$, as query dimensions, and the last two spatial dimensions, *i.e.*, $\mathbb{R}^{H_{\text{s}} \times W_{\text{s}}}$, as support dimensions hereafter.

Attentive squeeze layer (AS layer). The AS layer transforms a correlation tensor to another with a smaller support dimension via strided self-attention. The tensor is recast as a matrix with each element representing a support pattern. Given a correlation tensor $\mathbf{C} \in \mathbb{R}^{H_{\text{q}} \times W_{\text{q}} \times H_{\text{s}} \times W_{\text{s}} \times C_{\text{in}}}$ in a hypercorrelation pyramid, we start by reshaping the correlation tensor as a block matrix of size $H_{\text{q}} \times W_{\text{q}}$ with each element corresponding to a correlation tensor of $\mathbf{C}(\mathbf{x}_{\text{q}}) \in \mathbb{R}^{H_{\text{s}} \times W_{\text{s}} \times C_{\text{in}}}$ on the query position $\mathbf{x}_{\text{q}}$ such that

$$\mathbf{C}^{\text{block}} = \begin{bmatrix} \mathbf{C}((1,1)) & \dots & \mathbf{C}((1,W_{\text{q}})) \\ \vdots & \ddots & \vdots \\ \mathbf{C}((H_{\text{q}},1)) & \dots & \mathbf{C}((H_{\text{q}},W_{\text{q}})) \end{bmatrix}. \quad (9)$$

We call each element a *support correlation tensor*. The goal of an AS layer is to analyze the global context of each support correlation tensor and extract a correlational representation with a reduced support dimension while the query dimension is preserved: $\mathbb{R}^{H_{\text{q}} \times W_{\text{q}} \times H_{\text{s}} \times W_{\text{s}} \times C_{\text{in}}} \rightarrow \mathbb{R}^{H_{\text{q}} \times W_{\text{q}} \times H'_{\text{s}} \times W'_{\text{s}} \times C_{\text{out}}}$, where $H'_{\text{s}} \leq H_{\text{s}}$ and $W'_{\text{s}} \leq W_{\text{s}}$. To learn a holistic pattern of each support correlation, we adopt the global self-attention mechanism [75] for correlational feature transform. The self-attention weights are shared across all query positions and processed in parallel.

Let us denote a support correlation tensor on any query position $\mathbf{x}_{\text{q}}$ by $\mathbf{C}^{\text{s}} = \mathbf{C}^{\text{block}}(\mathbf{x}_{\text{q}})$ for notational brevity as all positions share the following computation. The self-attention computation starts by embedding a support correlation tensor $\mathbf{C}^{\text{s}}$ to a target [1] , key, value triplet: $\mathbf{T}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{H'_{\text{s}} \times W'_{\text{s}} \times C_{\text{hd}}}$, using three convolutions of which strides greater than or equal to one to govern the output size. The resultant target and key correlational representations, $\mathbf{T}$ and $\mathbf{K}$, are then used to compute an attention context. The attention context is computed as following matrix multiplication:

$$\mathbf{A} = \mathbf{T}\mathbf{K}^{\top} \in \mathbb{R}^{H'_{\text{s}} \times W'_{\text{s}} \times H'_{\text{s}} \times W'_{\text{s}}}. \quad (10)$$

Next, the attention context is normalized by softmax such that the votes on key foreground positions sum to one with masking attention by the support mask annotation $\mathbf{Y}_{\text{s}}$ if available to attend more on the foreground region:

$$\bar{\mathbf{A}}(\mathbf{p}_{\text{t}}, \mathbf{p}_{\text{k}}) = \frac{\exp\left(\mathbf{A}(\mathbf{p}_{\text{t}}, \mathbf{p}_{\text{k}})\mathbf{Y}_{\text{s}}(\mathbf{p}_{\text{k}})\right)}{\sum_{\mathbf{p}'_{\text{k}}} \exp\left(\mathbf{A}(\mathbf{p}_{\text{t}}, \mathbf{p}'_{\text{k}})\mathbf{Y}_{\text{s}}(\mathbf{p}'_{\text{k}})\right)},$$

$$\text{where } \mathbf{Y}_{\text{s}}(\mathbf{p}_{\text{k}}) = \begin{cases} 1 & \text{if } \mathbf{p}_{\text{k}} \in [H'_{\text{s}}] \times [W'_{\text{s}}] \text{ is foreground,} \\ -\infty & \text{otherwise.} \end{cases} \quad (11)$$

The masked attention context $\bar{\mathbf{A}}$ is then used to aggregate the value embedding $\mathbf{V}$:

$$\mathbf{C}^{\text{s}}_{\text{A}} = \bar{\mathbf{A}}\mathbf{V} \in \mathbb{R}^{H'_{\text{s}} \times W'_{\text{s}} \times C_{\text{hd}}}. \quad (12)$$

The attended representation is fed to an MLP layer, $\mathbf{W}_{\text{o}}$, and added to the input. In case the input and output dimensions mismatch, the input is optionally fed to a convolutional layer, $\mathbf{W}_{\text{I}}$. The addition is followed by an activation layer $\varphi(\cdot)$ consisting of a group normalization [85] and a ReLU activation [46]:

$$\mathbf{C}^{\text{s}}_{\text{o}} = \varphi(\mathbf{W}_{\text{o}}(\mathbf{C}^{\text{s}}_{\text{A}}) + \mathbf{W}_{\text{I}}(\mathbf{C}^{\text{s}})) \in \mathbb{R}^{H'_{\text{s}} \times W'_{\text{s}} \times C_{\text{out}}}. \quad (13)$$

The output is then fed to another MLP that concludes a unit operation of an AS layer:

$$\mathbf{C}^{\text{s}\prime} = \varphi(\mathbf{W}_{\text{FF}}(\mathbf{C}^{\text{s}}_{\text{o}}) + \mathbf{C}^{\text{s}}_{\text{o}}) \in \mathbb{R}^{H'_{\text{s}} \times W'_{\text{s}} \times C_{\text{out}}}, \quad (14)$$

which is embedded to the corresponding query position in the block matrix of Eq. (9). Note that the AS layer can be stacked to progressively reduce the size of support correlation tensor, $H'_{\text{s}} \times W'_{\text{s}}$, to a smaller size. The overall pipeline of AS layer is illustrated in the supplementary material.

Multi-layer fusion. The pyramid correlational representations are merged from the coarsest to the finest level by cascading a pair-wise operation of the following three steps: upsampling, addition, and non-linear transform. We first bi-linearly upsample the bottommost correlational representation to the query spatial dimension of its adjacent earlier one and then add the two representations to obtain a mixed one $\mathbf{C}^{\text{mix}}$. The mixed representation is fed to two sequential AS layers until it becomes a point feature of size $H'_{\text{s}} = W'_{\text{s}} = 1$, which is fed to the subsequent pyramidal fusion. The output from the earliest fusion layer is fed to a convolutional decoder, which consists of interleaved 2D convolution and bi-linear upsampling that map the C-dimensional channel to 2 (foreground and background) and the output spatial size to the input query image size. See Fig. 2 for the overall process of multi-layer fusion.

Class-wise foreground map computation. The K-shot output foreground activation maps are averaged to produce a mask prediction for each class. The averaged output map is normalized by softmax over the two channels of the binary segmentation map to obtain a foreground probability prediction $\mathbf{Y}^{(n)} \in \mathbb{R}^{H \times W}$.

[1] In this section, we adapt the term "target" to indicate the "query" embedding in the context of self-attention learning [11, 24, 53, 75, 79] to avoid homonymous confusion with the "query" image to be segmented.

	1-way 1-shot										2-way 1-shot									
	classification 0/1 exact ratio (%)					segmentation mIoU (%)					classification 0/1 exact ratio (%)					segmentation mIoU (%)				
method	5^0	5^1	5^2	5^3	avg.	5^0	5^1	5^2	5^3	avg.	5^0	5^1	5^2	5^3	avg.	5^0	5^1	5^2	5^3	avg.
PANet [78]	69.9	67.7	68.8	69.4	69.0	32.8	45.8	31.0	35.1	36.2	56.2	47.5	44.6	55.4	50.9	33.3	46.0	31.2	38.4	37.2
PFENet [73]	69.8	82.4	68.1	77.9	74.6	38.3	54.7	35.1	43.8	43.0	22.5	61.7	40.3	39.5	41.0	31.1	47.3	30.8	32.2	35.3
HSNet [44]	**86.6**	84.8	76.9	**86.3**	83.7	49.1	59.7	41.0	49.0	49.7	68.0	73.2	57.0	**70.9**	67.3	42.4	53.7	34.0	43.9	43.5
ASNet$_\text{w}$	86.4	86.3	70.9	84.5	82.0	10.8	20.2	13.1	16.1	15.0	**71.6**	72.4	46.4	68.0	64.6	11.4	20.8	12.5	15.9	15.1
ASNet	84.9	**89.6**	**79.0**	86.2	**84.9**	**51.7**	**61.5**	**43.3**	**52.8**	**52.3**	68.5	**76.2**	**58.6**	70.0	**68.3**	**48.5**	**58.3**	**36.3**	**48.3**	**47.8**

Table 1. Performance comparison of ASNet and others on FS-CS and Pascal-5^i [63]. All methods are trained and evaluated under the iFSL framework given strong labels, *i.e.*, class segmentation masks, except for ASNet$_\text{w}$ that is trained only with weak labels, *i.e.*, class tags.

	1-way 1-shot		2-way 1-shot	
method	ER	mIoU	ER	mIoU
PANet [78]	66.7	25.2	48.5	23.6
PFENet [73]	71.4	31.9	36.5	22.6
HSNet [44]	77.0	34.3	62.5	29.5
ASNet	**78.6**	**35.8**	**63.1**	**31.6**

Table 2. Performance comparison of ASNet and others on FS-CS and COCO-20^i [47].

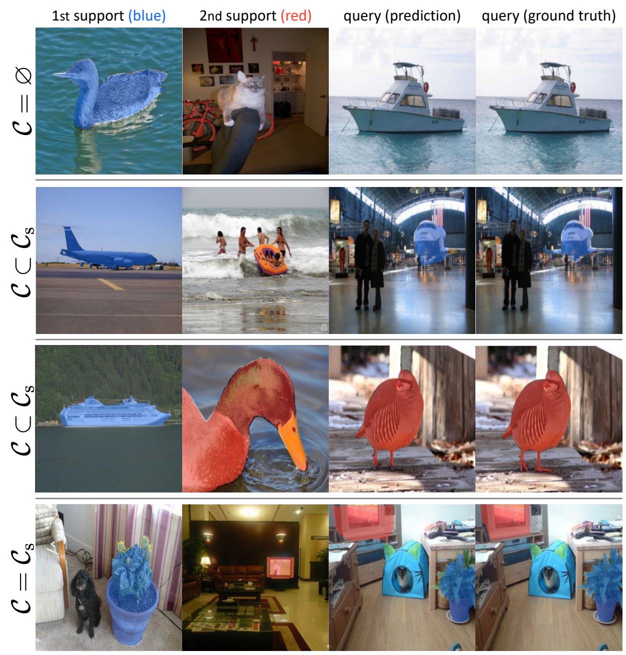

Figure 3. 2-way 1-shot segmentation results of ASNet on FS-CS. The examples cover all three cases of $\mathcal{C} = \varnothing$, $\mathcal{C} \subset \mathcal{C}_\text{s}$, and $\mathcal{C} = \mathcal{C}_\text{s}$. The images are resized to square shape for visualization.

6. Experiments

In this section we report our experimental results regarding the FS-CS task, the iFSL framework, as well as the ASNet after briefly describing implementation details and evaluation benchmarks. See the supplementary material for additional results, analyses, and experimental details.

6.1. Experimental setups

Experimental settings. We select ResNet50 and ResNet-101 [22] pretrained on ImageNet [60] as our backbone networks for a fair comparison with other methods and freeze the backbone during training as similarly as the previous work [44, 73]. We train models using Adam [27] optimizer with learning rate of 10^{-4} and 10^{-3} for the classification loss and the segmentation loss, respectively. We train all models with 1-way 1-shot training episodes and evaluate the models on arbitrary N-way K-shot episodes. For inferring class occurrences, we use a threshold $\delta = 0.5$. All the AS layers are implemented as multi-head attention with 8 heads. The number of correlation pyramid is set to $P = 3$.

Dataset. For the new task of FS-CS, we construct a benchmark adopting the images and splits from the two widely-used FS-S datasets, Pascal-5^i [13, 63] and COCO-20^i [36, 47], which are also suitable for multi-label classification [83]. Within each fold, we construct an episode by randomly sampling a query and an N-way K-shot support set that annotates the query with N-way class labels and an $(N{+}1)$-way segmentation mask in the context of the support set. For the FS-S task, we also use Pascal-5^i and COCO-20^i following the same data splits as [63] and [47], respectively.

Evaluation. Each dataset is split into four mutually disjoint class sets and cross-validated. For multi-label classification evaluation metrics, we use the 0/1 exact ratio $\text{ER} = \mathbb{1}[\mathbf{y}_\text{gt} = \mathbf{y}_\text{C}]$ [12]. In the supplementary material, we also report the results in accuracy $\text{acc} = \frac{1}{N}\sum_n \mathbb{1}[\mathbf{y}_\text{gt}^{(n)} = \mathbf{y}_\text{C}^{(n)}]$. For segmentation, we use mean IoU $\text{mIoU} = \frac{1}{C}\sum_c \text{IoU}_c$ [63, 78], where IoU_c denotes an IoU value of c_th class.

6.2. Experimental evaluation of iFSL on FS-CS

In this subsection, we investigate the iFSL learning framework on the FS-CS task. All ablation studies are conducted using ResNet50 on Pascal-i^5 and evaluated in 1-way 1-shot setup unless specified otherwise. Note that it is difficult to present a fair and direct comparison between the conventional FS-C and our few-shot classification task since FS-C is always evaluated on single-label classification benchmarks [2, 32, 57, 74, 76], whereas our task is evaluated on multi-label benchmarks [13, 36], which are irreducible to a single-label one in general.

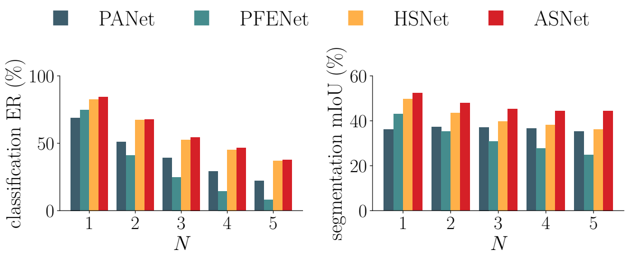

Figure 4. N-way 1-shot FS-CS performance comparison of four methods by varying N from 1 to 5.

Effectiveness of iFSL on FS-CS. We validate the iFSL framework on FS-CS and also compare the performance of ASNet with those of three recent state-of-the-art methods, PANet [78], PFENet [73], and HSNet [44], which are originally proposed for the conventional FS-S task; all the models are trained by iFSL for a fair comparison. Note that we exclude the background merging step (Eqs. 3 and 4) for PANet as its own pipeline produces a multi-class output including background. Tables 1 and 2 validate the iFSL framework on the FS-CS task quantitatively, where our ASNet surpasses other methods on both 1-way and 2-way setups in terms of few-shot classification as well as the segmentation performance. The 2-way segmentation results are also qualitatively demonstrated in Fig. 3 visualizing exhaustive inclusion relations between a query class set $\mathcal{C}$ and a target (support) class set $\mathcal{C}_\text{s}$ in a 2-way setup.

Weakly-supervised iFSL. The iFSL framework is versatile across the level of supervision: weak labels (class tags) or strong labels (segmentation masks). Assuming weak labels are available but strong labels are not, ASNet is trainable with the classification learning objective of iFSL (Eq. 6) and its results are presented as ASNet$_\text{w}$ in Table 1. ASNet$_\text{w}$ performs on par with ASNet in terms of classification ER (82.0% *vs.* 84.9% on 1-way 1-shot), but performs ineffectively on the segmentation task (15.0% *vs.* 52.3% on 1-way 1-shot). The result implies that the class tag labels are sufficient for a model to recognize the class occurrences, but are weak to endorse model's precise spatial recognition ability.

Multi-class scalability of FS-CS. In addition, FS-CS is extensible to a multi-class problem with arbitrary numbers of classes, while FS-S is not as flexible as FS-CS in the wild. Figure 4 compares the FS-CS performances of four methods by varying the N-way classes from one to five, where the other experimental setup follows the same one as in Table 1. Our ASNet shows consistently better performances than other methods on FS-CS in varying number of classes.

Robustness of FS-CS against task transfer. We evaluate the transferability between FS-CS, FS-C, and FS-S by training a model on one task and evaluating it on the other task. The results are compared in Fig. 5 in which 'FS-S $\rightarrow$ FS-CS' represents the result where the model trained on the

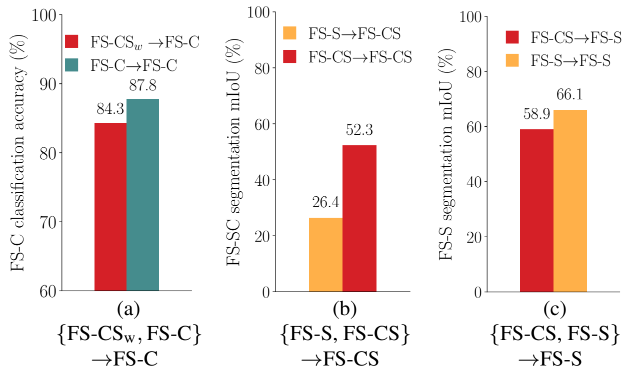

Figure 5. Results of task transfer. A $\rightarrow$ B denotes a model trained on task A and evaluated on task B. FS-CS$_\text{w}$ denotes FS-CS with weak labels. **(a)**: Exclusive 2-way 1-shot classification accuracy of FS-C or FS-CS$_\text{w}$ learners on FS-C. **(b)**: 1-way 1-shot segmentation mIoU of FS-S or FS-CS learners on FS-CS. **(c)**: 1-way 1-shot segmentation mIoU of FS-S or FS-CS learners on FS-S.

FS-S task (with the guarantee of support class presence) is evaluated on the FS-CS setup. To construct training and validation splits for FS-C or FS-S, we sample episodes that satisfy the constraint of support class occurrences [2]. For training FS-C models, we use the class tag supervision only. All the other settings are fixed the same, *e.g.*, we use ASNet with ResNet50 and Pascal-i^5.

The results show that FS-CS learners, *i.e.*, models trained on FS-CS, are transferable to the two conventional few-shot learning tasks and yet overcome their shortcomings. The transferability between few-shot classification tasks, *i.e.*, FS-C and FS-CS$_\text{w}$, is presented in Fig. 5 (a). On this setup, the FS-CS$_\text{w}$ learner is evaluated by predicting a higher class response between the two classes, although it is trained using the multi-label classification objective. The FS-CS learner closely competes with the FS-C learner on FS-C in terms of classification accuracy. In contrast, the task transfer between segmentation tasks, FS-S and FS-CS, results in asymmetric outcomes as shown in Fig. 5 (b) and (c). The FS-CS learner shows relatively small performance drop on FS-S, however, the FS-S learner suffers a severe performance drop on FS-CS. Qualitative examples in Fig. 1 demonstrate that the FS-S learner predicts a vast number of false-positive pixels and results in poor performances. In contrast, the FS-CS learner successfully distinguishes the region of interest by analyzing the semantic relevance of the query objects between the support set.

6.3. Comparison with recent FS-S methods on FS-S

Tables 3 and 4 compare the results of the recent few-shot semantic segmentation methods and ASNet on the conven-

[2] We sample 2-way 1-shot episodes having a single positive class for training on FS-C or evaluating on FS-C. We collect 1-way 1-shot episodes sampled from the same class for training on FS-S or evaluating on FS-S.

	method	1-way 1-shot 5^0	5^1	5^2	5^3	mIoU	FBIoU	1-way 5-shot 5^0	5^1	5^2	5^3	mIoU	FBIoU	# learn. params.
R50	CANet [98]	52.5	65.9	51.3	51.9	55.4	66.2	55.5	67.8	51.9	53.2	57.1	69.6	-
	PPNet [40]	47.8	58.8	53.8	45.6	51.5	69.2	58.4	67.8	64.9	56.7	62.0	75.8	23.5 M
	PFENet [73]	61.7	69.5	55.4	56.3	60.8	73.3	63.1	70.7	55.8	57.9	61.9	73.9	31.5 M
	SAGNN [87]	64.7	69.6	57.0	57.2	62.1	73.2	64.9	70.0	57.0	59.3	62.8	73.3	-
	MMNet [86]	62.7	70.2	57.3	57.0	61.8	-	62.2	71.5	57.5	62.4	63.4	-	10.4 M
	CMN [88]	64.3	70.0	57.4	59.4	62.8	72.3	65.8	70.4	57.6	60.8	63.7	72.8	-
	MLC [90]	59.2	71.2	**65.6**	52.5	62.1	-	63.5	71.6	**71.2**	58.1	66.1	-	8.7 M
	HSNet [44]	64.3	70.7	60.3	60.5	64.0	76.7	70.3	73.2	67.4	**67.1**	69.5	**80.6**	2.6 M
	ASNet	**68.9**	**71.7**	61.1	**62.7**	**66.1**	**77.7**	**72.6**	**74.3**	65.3	**67.1**	**70.8**	80.4	**1.3 M**

Table 3. FS-S results on 1-way 1-shot and 1-way 5-shot setups on Pascal-5^i [63] using ResNet50 [22] (R50).

	method	1-way 1-shot mIoU	FBIoU	1-way 5-shot mIoU	FBIoU	# learn. params.
R50	RPMM [89]	30.6	-	35.5	-	38.6 M
	RePRI [3]	34.0	-	42.1	-	-
	MMNet [86]	37.5	-	38.2	-	10.4 M
	MLC [90]	33.9	-	40.6	-	8.7 M
	CMN [88]	39.3	61.7	43.1	63.3	-
	HSNet [44]	39.2	68.2	46.9	70.7	2.6 M
	ASNet	**42.2**	**68.8**	**47.9**	**71.6**	**1.3 M**

Table 4. FS-S results on 1-way 1-shot and 1-way 5-shot setups on COCO-20^i [47].

method	ER	mIoU
(a) global → local	83.9	44.6
(b) w/o masked attention	83.8	50.8
(c) w/o multi-layer fusion	83.1	51.6
ASNet	84.9	52.3

Table 5. Ablation study of the AS layer on 1-way 1-shot on Pascal-5^i [63] using ResNet50 [22].

tional FS-S task. All model performances in the tables are taken from corresponding papers, and the numbers of learnable parameters are either taken from papers or counted from their official sources of implementation. For a fair comparison with each other, some methods that incorporate extra unlabeled images [40, 90] are reported as their model performances measured in the absence of the extra data. Note that ASNet in Tables 3 and 4 is trained and evaluated following the FS-S setup, not the proposed FS-CS one.

The results verify that ASNet outperforms the existing methods including the most recent ones [86, 88, 90]. Especially, the methods that cast few-shot segmentation as the task of correlation feature transform, ASNet and HSNet [44], outperform other visual feature transform methods, indicating that learning correlations is beneficial for both FS-CS and FS-S. Note that ASNet is the most lightweight among others as ASNet processes correlation features that have smaller channel dimensions, *e.g.*, at most 128, than visual features, *e.g.*, at most 2048 in ResNet50.

6.4. Analyses on the model architecture

We perform ablation studies on the model architecture to reveal the benefit of each component. We replace the global self-attention in the ASNet layer with the local self-attention [53] to see the effect of the global self-attention (Table 5a). The local self-attention variant is compatible with the global ASNet in terms of the classification exact ratio but degrades the segmentation mIoU significantly, signifying the importance of the learning the global context of feature correlations. Next, we ablate the attention masking in Eq. (11), which verifies that the attention masking prior is effective (Table 5b). Lastly, we replace the multi-layer fusion path with spatial average pooling over the support dimensions followed by element-wise addition (Table 5c), and the result indicates that it is crucial to fuse outputs from the multi-layer correlations to precisely estimate class occurrence and segmentation masks.

7. Discussion

We have introduced the integrative task of few-shot classification and segmentation (FS-CS) that generalizes two existing few-shot learning problems. Our proposed integrative few-shot learning (iFSL) framework is shown to be effective on FS-CS, in addition, our proposed attentive squeeze network (ASNet) outperforms recent state-of-the-art methods on both FS-CS and FS-S. The iFSL design allows a model to learn either with weak or strong labels, that being said, learning our method with weak labels achieves low segmentation performances. This result opens a future direction of effectively boosting the segmentation performance leveraging weak labels in the absence of strong labels for FS-CS.

Acknowledgements. This work was supported by Samsung Advanced Institute of Technology (SAIT) and also by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD).

References

[1] Kelsey Allen, Evan Shelhamer, Hanul Shin, and Joshua Tenenbaum. Infinite mixture prototypes for few-shot learning. In *Proc. International Conference on Machine Learning (ICML)*, 2019. 2

[2] Luca Bertinetto, Joao F Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *Proc. International Conference on Learning Representations (ICLR)*, 2018. 6

[3] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 8

[4] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern recognition*, 2004. 2

[5] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proc. European Conference on Computer Vision (ECCV)*, 2018. 2

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. 1

[7] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[8] Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[9] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *International Conference on Learning Representations*, 2019. 2

[10] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *Proc. British Machine Vision Conference (BMVC)*, 2018. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. International Conference on Learning Representations (ICLR)*, 2021. 2, 5

[12] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 647–657, 2019. 2, 6

[13] 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 (IJCV)*, 2010. 6

[14] Geli Fei and Bing Liu. Breaking the closed world assumption in text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016. 2

[15] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2006. 1

[16] Michael Fink. Object classification from a single example utilizing class relevance metrics. *Advances in Neural Information Processing Systems (NeurIPS)*, 2005. 1

[17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. International Conference on Machine Learning (ICML)*, 2017. 2, 3

[18] Siddhartha Gairola, Mayur Hemani, Ayush Chopra, , and Balaji Krishnamurthy. Simpropnet: Improved similarity propagation for few-shot image segmentation. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2020. 2

[19] Dan Andrei Ganea, Bas Boom, and Ronald Poppe. Incremental few-shot instance segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[20] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[21] Guangxing Han, Yicheng He, Shiyuan Huang, Jiawei Ma, and Shih-Fu Chang. Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 4, 6, 8

[23] Ruibing Hou, Hong Chang, MA Bingpeng, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[24] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 5

[25] Tao Hu, Pengwan Yang, Chiliang Zhang, Gang Yu, Yadong Mu, and Cees G. M. Snoek. Attention-based multi-context guiding for few-shot semantic segmentation. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 1

[26] Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho. Relational embedding for few-shot classification. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proc. International Conference on Learning Representations (ICLR)*, 2015. 6

[28] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *International Conference on Machine Learning (ICML) Workshop on Deep Learning*, 2015. 1, 2

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012. 1

[30] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 2015. 1

[31] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[32] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[33] Li-Jia Li, Richard Socher, and Li Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 1

[34] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4

[35] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. European Conference on Computer Vision (ECCV)*, 2014. 6

[37] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2

[38] Binghao Liu, Yao Ding, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Anti-aliasing semantic reconstruction for few-shot semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[39] Bo Liu, Hao Kang, Haoxiang Li, Gang Hua, and Nuno Vasconcelos. Few-shot open-set recognition using meta-learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[40] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2, 8

[41] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[42] Andrew Kachites McCallum. Multi-label text classification with a mixture model trained by em. In *AAAI 99 workshop on text learning*, 1999. 2

[43] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*. 1989. 2

[44] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 4, 6, 7, 8

[45] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[46] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning (ICML)*, 2010. 4, 5

[47] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 6, 8

[48] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[49] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[50] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[51] Limeng Qiao, Yemin Shi, Jia Li, Yaowei Wang, Tiejun Huang, and Yonghong Tian. Transductive episodic-wise adaptive metric for few-shot learning. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[52] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alexei Efros, and Sergey Levine. Conditional networks for few-shot semantic segmentation. 2018. 1, 3

[53] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 5, 8

[54] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[55] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *Proc. International Conference on Learning Representations (ICLR)*, 2017. 1, 3

[56] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[57] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. In *Proc. International Conference on Learning Representations (ICLR)*, 2018. 6

[58] Pau Rodríguez, Issam Laradji, Alexandre Drouin, and Alexandre Lacoste. Embedding propagation: Smoother manifold for few-shot classification. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2

[59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 1

[60] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 6

[61] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2018. 2

[62] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2012. 2

[63] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *Proc. British Machine Vision Conference (BMVC)*, 2017. 1, 3, 6, 8

[64] Mennatullah Siam, Naren Doraiswamy, Boris N Oreshkin, Hengshuai Yao, and Martin Jagersand. Weakly supervised few-shot object segmentation using co-attention with visual and semantic embeddings. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2020. 2

[65] Mennatullah Siam, Boris N. Oreshkin, and Martin Jagersand. Amp: Adaptive masked proxies for few-shot segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[66] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Proc. International Conference on Learning Representations (ICLR)*, 2015. 1

[67] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 3

[68] Guolei Sun, Yun Liu, Jingyun Liang, and Luc Van Gool. Boosting few-shot semantic segmentation with transformers. *arXiv preprint arXiv:2108.02266*, 2021. 2

[69] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[70] Pinzhuo Tian, Zhangkai Wu, Lei Qi, Lei Wang, Yinghuan Shi, and Yang Gao. Differentiable meta-learning model for few-shot semantic segmentation. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 2

[71] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2

[72] Zhuotao Tian, Xin Lai, Li Jiang, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. *arXiv preprint arXiv:2010.05210*, 2020. 2

[73] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 2, 6, 7, 8

[74] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *Proc. International Conference on Learning Representations (ICLR)*, 2020. 6

[75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 5

[76] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 1, 2, 3, 6

[77] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2

[78] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 6, 7

[79] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 5

[80] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[81] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *Proc. International Conference on Machine Learning (ICML)*, 2020. 2

[82] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 2020. 1

[83] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017. 6

[84] Yan Wu and Yiannis Demiris. Towards one shot learning by imitation for humanoid robots. In *2010 IEEE International Conference on Robotics and Automation*, 2010. 1

[85] Yuxin Wu and Kaiming He. Group normalization. In *Proc. European Conference on Computer Vision (ECCV)*, 2018. 5

[86] Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. Learning meta-class memory for few-shot semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 8

[87] Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. Scale-aware graph neural network for few-shot semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 8

[88] Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao. Few-shot semantic segmentation with cyclic memory network. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 8

[89] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Ye Qixiang. Prototype mixture models for few-shot semantic segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2, 8

[90] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. Mining latent classes for few-shot segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 8

[91] Xianghui Yang, Bairun Wang, Kaige Chen, Xinchi Zhou, Shuai Yi, Wanli Ouyang, and Luping Zhou. Brinet: Towards bridging the intra-class and inter-class gaps in one-shot segmentation. 2020. 2

[92] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1

[93] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[94] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[95] Bingfeng Zhang, Jimin Xiao, and Terry Qin. Self-guided and cross-guided learning for few-shot segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[96] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[97] 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 *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[98] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 8

[99] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Few-shot 3d point cloud semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[100] Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui, and Ling Shao. Collaborative learning of semi-supervised segmentation and classification for medical images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1