# HyperClass: A Hypernetwork for Few Shot One-Class Classification and Open Set Recognition

Boaz Lerner Nir Darshan Rami Ben-Ari
OriginAI, Israel
{boazl, nir, ramib} @originai.co

## Abstract

*Achieving generalization across unobserved and diverse distributions with limited data is a challenge for modern machine learning methods. Few-shot one-class classification offers the capability to discriminate between a desired target class and instances from unseen or unknown categories, even with only a limited number of target samples at hand. In this paper we introduce a new approach for few-shot one-class classification and open set recognition. Our scheme is based on a hyper-network, that generates a binary classifier, learned by Meta-learning. We demonstrate the advantage of our model in attaining SoTA results in few-shot one-class classification while reaching competitive results in few-Shot open-set recognition task.*

## 1. Introduction

*One-class classification* (OCC), focuses on identifying instances of a specific target class, utilizing labeled samples exclusively from that class. At test, OCC models act as binary classifier discriminating between the target (also known as *positive*) and other classes (also known as *negative*) that can be combined of numerous unseen classes (11). Therefore OCC exhibits a case where a classifier should cope with a multi-modal and out-of-distribution set. Applications of OCC include outlier-detection *e.g.* (3; 15) and anomaly detection (2).

Extending the positive class from one category to multiple (known) classes, introduces the well known problem of *open set recognition* (OSR), a sub-task of Open Set Classification problem (1; 6). In open set classification one has to deal with unseen classes, existing as distractors. The task requires the additional binary classification sub-task of OSR for discriminating between the known *in-set* classes and unknown *out-of-set* and out-of-distribution distractors.

Few-shot OCC (FSOCC) (5; 12; 13) attempts to learn the one-class classifier with only few positive samples. In Few-Shot Open-Set Recognition (FSOR) (9; 10; 14), one has to deal with unseen classes, effectively generalizing to an unseen distribution, having only few labeled samples from the in-set classes.

Only few works has previously addressed the FSOCC task (5; 12; 13). Kruspe (13) suggested a prototypical network (18) with additional "null" (garbage class), which does not need to be labeled. (5) presented a method to modify the episodic data sampling in the model-agnostic meta-learning (MAML) algorithm to learn a new model initialization for FSOCC. Kozerawski and Turk (12) suggested the Meta Binary Cross-Entropy (Meta-BCE), which learns a separate feature representation for one-class classification, and One-Class Meta-Learning (OCML), which starts with a transfer learning module. MAML methods has gained popularity in FSL. Yet, the huge computational graphs, primarily stemming by the end-to-end train of the backbone (feature extractor), frequently result in overfitting and is challenging to train in practice. A common remedy, is choosing a shallow backbone *e.g.* Conv-4 as used in (5; 12), compromising features quality and eventually performance.

In this study we suggest a model that incorporates two main ingredients. First, we suggest a MAML strategy, built on a frozen feature extractor, with task episode sampling that mimics the FSOCC problem characteristics. The second aspect refers to the design of our model and its training strategy (see Fig. 1 for a schematic view of the architecture). We decompose the standard linear classifier typically used in FSL to two components. A "global" vector $v$ that represents a prior over a general task, and an adaptation matrix $P$, adapting the few-shot classifier to a specific class or in-set categories, both learned in MAML fashion. Overall, our contribution is three-fold: 1) We suggest a novel learning approach for FSOCC and FSOR. Our approach offers computationally efficient training, which can accommodate a variety of deep backbones. 2) Our approach can be easily extended to transductive learning. 3) We evaluate our model on miniImagenet and tieredImageNet and present SoTA performance on FSOCC and achieve competitive results on FSOR task.

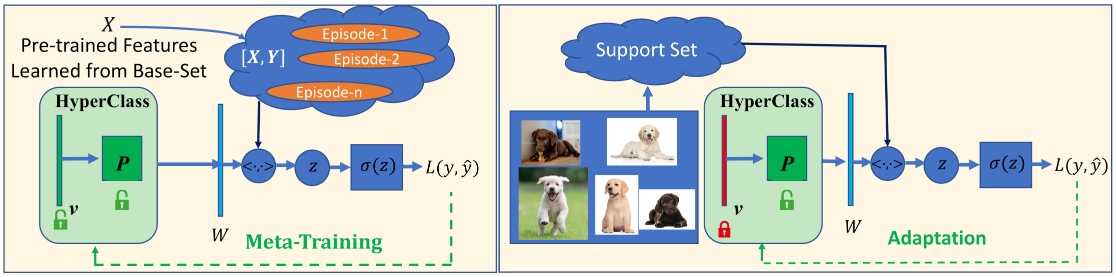


Figure 1: Our HyperClass trainig scheme. **Left:** Meta Training: Using pre-trained fixed features $X$, we learn a task-agnostic classifier $v$ and a projection head $P$, via MAML. The model outputs a linear classifier $W$, $z = W^T x$, $\sigma$ is the sigmoid function and $L$ is BCE loss. **Right:** Test Adaptation: Given the few labeled samples in the support set (from one/multiple classes), HyperClass runs several adaptation steps to reach a final classifier for the specific in-set categories (see details in Sec. 2.1).

## 2. Method

In few-shot image classification, the evaluation called *meta-testing* is conducted on randomly sampled $N$-way novel classes with $K$-shot samples per-class. The $K$-shot samples are used as *support set* and employed for adaptation/training, whereas different samples from those $N$ novel classes are employed as *query set* and used for testing.

Our model is a small hyper-network, dubbed HyperClass that outputs a linear binary classifier. HyperClass seeks to benefit from the two main approaches of Few-Shot Learning (FSL): Transfer Learning and Meta Learning. We therefore divide the training into two stages. The first one aims to learn discriminative features. Then, in the second stage, we apply a MAML approach on top of the pre-computed features. In contrast to common metric learning methods and MAML that adapt the embeddings in meta-train and/or in meta-test, we keep the embedding fixed, and train an adaptation module specialized to the specific embeddings. This adaptation module learns a task-agnostic classifier and a task-specific projection head adapting the classifier to a new task, namely the new target class in OCC or in-set classification in OSR. Figure 1 illustrates our training scheme and the associated components.

**Feature Extraction:** The goal of transfer learning is to learn a transferable embedding model $f_\phi$ which generalizes to a new task. To this end, we train a feature extractor on a standard multi-class task similar to (19). This is depicted in Fig. 1(a).

**Adaptation Model:** We start by describing our meta-training scheme. In our setting an image $I$ is mapped to a fixed representation $x = f_\phi(I) \in \mathbb{R}^d$ ($d$ - feature dimension). A straight forward and effective approach suggests learning a Logistic Regression (LR) classifier over the learned representations with the few samples from the support set, as in (19). The LR parameters $\theta_l = \{W_l, b_l\}$ then include a weight term $W_l \in \mathbb{R}^{N \times d}$ (N-number of classes) and a bias term $b_l \in \mathbb{R}^d$ obtained by:

$$\theta_l = \arg\min_{W_l, b_l} \sum_{i=1}^{n} \mathcal{L}^{bce}(W_l x_i^s + b_l, y_i^s) + \mathcal{R}(W_l, b_l), \quad (1)$$

where $D^s = \{(x_i^s, y_i^s)\}_{i=1}^n$ are the features and labels of the support set samples, $n$ denotes the number of support samples, and $\mathcal{L}^{bce}$ is the binary cross-entropy (BCE) loss. Finally, $\mathcal{R}(\cdot)$ is a regularization function.

We are also interested in a linear classifier due to scarcity of labeled samples and computational efficiency. However, we wish to leverage the information that variety of tasks has in common, and not to rely only on the task at hand. To this end, we learn a *prior*, represented by a *global* classifier and a good initialization for a projection head, adapting that classifier to a new task. The idea is to create an adapted classifier for each task with few samples and few gradient steps. This is performed with MAML optimization. We focus on binary classification, hence the classifier weights are reduced to $W \in \mathbb{R}^d$. More specifically, we introduce $v \in \mathbb{R}^d$ - a global *task-agnostic* classifier that is *projected* (transformed) into a subspace for each specific (local) task by:

$$W = Pv + b \quad (2)$$

where $P \in \mathbb{R}^{d \times d}$ is the projection head and $b \in \mathbb{R}^d$ is a bias term. The prediction is then $\hat{y} = \sigma(W^T x)$, with $\sigma$ standing for the Sigmoid function. For a schematic illustration of this part see Fig. 1(b)). More specifically, our trainable parameters $\theta = \{v, P, b\}$ are updated with a few gradient steps on the support set of a given task $j$, with the following rule:

$$\theta_j^{inner} = \theta - \nabla_\theta \mathcal{L}^s(\theta, D^s) + \mathcal{R}(\theta), \quad (3)$$

where $\mathcal{L}^s(\theta, D^s)$ refers to the BCE loss computed on predictions $\{\hat{y}_i^s\}_{i=1}^n$ and labels $\{y_i^s\}_{i=1}^n$. Then, in the outer loop we apply the adapted classifier on the query set for updating the global parameters, considering all inner loop updates, again with BCE loss $\mathcal{L}^q$, as follows:

$$\theta^{outer} = \theta - \frac{1}{T}\sum_{j=1}^{T} \nabla_\theta \mathcal{L}^q(\theta_j^{inner}, D^q), \qquad (4)$$

where $T$ is the number of meta-batch tasks and $D^q$ is the query set.

An outline of our meta-training pipeline is illustrated in Fig. 1. The figure describes the meta-training and adaptation stages. In meta-training the model is trained over various task-episodes with both $v$ and $P$. In each cycle of FSOCC/FSOR, the vector $v$ remains fixed, while $P$ is updated (initialized from the meta-training process).

### 2.1. Training strategy for different use cases:

In both FSOCC and FSOR, we simulate the meta-testing scenario with binary, and open-set tasks, while we impose the support set *e.g.* for FSOCC to have few samples from *one* class. The query set remains the same, including both positive and negative samples. Note that in FSOCC task, there is a trivial solution (mode collapse) due to BCE loss trained on one class. We avoid this solution by taking the following actions, including negative samples in the query set, applying just a few adaptation steps and $L_2$ regularization. We also report results with transductive variant of our method which makes use of the query set features without the labels, during adaptation on the FSOCC task. In this setting the loss is a weighted combination of the BCE loss on the labeled set together with entropy minimization (7) on the unlabeled query set. To handle FSOR, with multiple classes in meta-testing, we train a distinct projection head for every support class, utilizing solely samples derived from that specific class. For in-class probability, we take the max score over the trained classifiers. The meta-training model is the same as in FSOCC.

## 3. Experiments

We evaluate our method comparing it to SoTA and strong baselines on FSOCC and FSOR. In all benchmarks, we start by training a feature extractor on the base-classes in a standard multi-class classification task similar to (19).

### 3.1. Few-Shot One Class Classification

**Datasets:** For FSOCC, we conducted experiments on two common benchmarks for few-shot learning: miniImagenet (20), and tieredImageNet (17), both are subsets of the ImageNet dataset (4). The miniImagenet dataset consists of 100 categories from ImageNet with 84×84 RGB images, 600 per category. We follow the standard split of training/-validation/testing with 64/16/20 classes (16). The tieredImageNet dataset has 608 visual categories, split to 351/97/60 training/validation/testing categories, with a total of 450K 84x84 RGB images as in Ye et. al. (17). For evaluation, we follow the same protocol as in (5; 12), and report the average performance with a 95% confidence interval on 10K meta-testing tasks, where each task contains a 1-way K-shot support set and a query set with 15 positive and 15 negative samples. We use three evaluation metrics: Accuracy, F1-score, and AUROC (Area Under ROC curve), with thresholds for accuracy and F1-score in our experiments obtained according to the validation set.

**Compared Methods:** Here we compare our method to previous SoTA (12; 5) together with two strong baselines that we proposed: One-class SVM (OC-SVM) applied on the pre-trained features, and **Proto**, which takes the mean feature vector of the support set as the classifier.

### 3.2. Few Shot Open Set Recognition

Few-shot Open set recognition suggests a benchmark that divides the problem into two subtasks, *negative detection* (*i.e*. out-of-set classification) and a multi-class closed set classification. The negative detection is a binary classification task separating the "unseen" out-of-set class samples from the in-set samples belonging to categories in the support set. In literature, there is a separate performance report for each task.

Since our HyperClass generates a binary classifier, we further implemented a naïve generalization to the FSOR negative-detection task. We use this benchmark to showcase the capability of our classifier in this unique use-case. In contrast to FSOCC, in this scenario the positive (in) set is composed of $N$-way categories, where commonly $N = 5$. We therefore train a different projection head for each support class (with the provided labeled examples). Given a query sample, we then take the max probability over the trained classifiers as in-set probability. Note that we leave out the closed-set recognition task, since these two tasks are separated in the benchmark and any standard few-shot multi-class classification can be used *e.g.* (8; 19) to address it.

**Datasets:** Following FSOR test protocol, we evaluate our method on miniImagenet (20) and tieredImageNet (17) benchmarks with the same train/val/test splits as in FSOCC, and use AUROC measures for the negative detection task. In this scenario, a meta-testing episode is constructed by sampling a regular $N$-way $K$-shot support set, whereas the query set consists of 15 samples from each support class and additional 75 samples from 5 unknown classes (15 each), sampled randomly from the remaining non-support classes. This leads to a total of 75 known queries and 75 unknown queries for a 5-way episode.

**Compared Methods:** We compare ourselves to a) SoTA methods (9; 10; 14) b) Baselines generated from One-Class SVM and Proto.

| | | 1-shot | | | 5-shot | | |
|---|---|---|---|---|---|---|---|
| miniImagenet | | | | | | | |
| Model | Backbone | Acc | F1-Score | AUROC | Acc | F1-Score | AUROC |
| Upper Bound | ResNet12 | 86.15 ± 0.1 | 88.02 ± 0.16 | 94.45 ± 0.22 | 86.15 ± 0.1 | 88.02 ± 0.16 | 94.45 ± 0.22 |
| Meta BCE (12) | Conv-64 | 57.57 ± 0.78 | 28.3 ± 1.9 | 76.4 ± 1.6 | 77.38 ± 0.67 | 79.3 ± 0.5 | 85.0 ± 0.7 |
| OCML(12) | Conv-64 | 68.05 ± 0.99 | 62.0 ± 1.4 | 75.9 ± 1.2 | 74.74 ± 0.88 | 69.3 ± 1.3 | 85.4 ± 0.9 |
| OC-MAML(5) | Conv-4* | 69.1 | - | - | 76.2 | - | - |
| OC-SVM | Conv-4 | 50.12 ± 0.02 | 66.72 ± 0.01 | 77.44 ± 0.29 | 75.16 ± 0.21 | 76.79 ± 0.20 | 84.68 ± 0.20 |
| ProtoNet | Conv-4 | 66.73 ±0.22 | 71.27±0.24 | 77.48±0.29 | 69.89 ± 0.18 | 75.23 ± 0.17 | 84.48 ± 0.21 |
| HC (Ours) | Conv-4 | 68.03 ± 0.24 | 69.05 ± 0.33 | 78.45 ± 0.29 | 77.69 ± 0.19 | 76.84 ± 0.23 | 86.83 ± 0.18 |
| OC-SVM | ResNet12 | 75.14 ± 0.24 | 72.74 ± 0.35 | 85.23 ± 0.23 | 81.36 ± 0.19 | 82.59 ± 0.17 | 92.10 ± 0.12 |
| ProtoNet | ResNet12 | 76.12 ± 0.24 | 72.02 ± 0.38 | 85.23 ± 0.23 | 82.20 ± 0.19 | 79.78 ± 0.28 | 92.11 ± 0.12 |
| HC (Ours) | ResNet12 | **76.66 ± 0.23** | **76.22 ± 0.31** | **85.72 ± 0.23** | **84.93 ± 0.15** | **85.48 ± 0.15** | **92.98 ± 0.11** |
| HC - Transuctive (Ours) | ResNet12 | **78.74 ± 0.24** | **77.95 ± 0.23** | **89.59 ± 0.15** | **85.42 ± 0.18** | **86.12 ± 0.13** | **93.23 ± 0.11** |
| tieredImageNet | | | | | | | |
| Upper Bound | ResNet12 | 86.33 ± 0.17 | 88.67 ± 0.14 | 95.49 ± 0.27 | 86.33 ± 0.17 | 88.67 ± 0.14 | 95.49 ± 0.27 |
| Meta BCE (12) | Conv-64 | 55.87 ± 0.38 | 20.8 ± 1.0 | 75.5 ± 0.9 | 75.62 ± 0.65 | 77.4 ± 0.5 | 83.1 ± 0.7 |
| OCML (12) | Conv-64 | 72.13 ± 0.33 | 72.4 ± 0.3 | 80.0 ± 0.4 | 78.89 ± 0.67 | 80.3 ± 0.6 | 87.8 ± 0.7 |
| OC-SVM | Conv-4 | 68.66 ± 0.24 | 72.18 ± 0.25 | 79.63 ± 0.28 | 77.02 ± 0.19 | 74.22 ± 0.25 | 87.17 ± 0.18 |
| ProtoNet | Conv-4 | 71.13 ± 0.24 | 71.95 ± 0.31 | 79.66 ± 0.28 | 77.66 ± 0.19 | 79.08 ± 0.21 | 87.10 ± 0.18 |
| HC (Ours) | Conv-4 | 70.98 ± 0.25 | 70.76 ± 0.34 | 80.48 ± 0.28 | 79.53 ± 0.19 | 81.28 ± 0.17 | 88.62 ± 0.17 |
| OC-SVM | ResNet12 | 78.39 ± 0.25 | 75.52 ± 0.38 | 87.94 ± 0.22 | 84.50 ± 0.17 | 84.70 ± 0.18 | 93.59 ± 0.12 |
| ProtoNet | ResNet12 | 76.29 ± 0.26 | 68.58 ± 0.46 | 87.95 ± 0.22 | 80.24 ± 0.25 | 73.45 ± 0.45 | 93.49 ± 0.13 |
| HC (Ours) | ResNet12 | **79.01 ± 0.24** | **76.66 ± 0.36** | **88.40 ± 0.22** | **84.77 ± 0.18** | **86.58 ± 0.14** | **94.11 ± 0.11** |
| HC - Transductive (Ours) | ResNet12 | **80.07 ± 0.22** | **78.93 ± 0.25** | **91.16 ± 0.15** | **84.93 ± 0.19** | **87.12 ± 0.12** | **94.32 ± 0.13** |

Table 1: Few-Shot One-Class Classification results on **miniImagenet** and **tieredImageNet** for various methods and backbones, [1] (12). Performance measures correspond to confidence interval of 95% over 10K episodes. OC-MAML reports only accuracy measure. Note that Conv-64 and Conv-4* are larger/optimized architectures w.r.t our simple Conv-4, commonly used in MAML methods. We also report the results with our transductive version marked in blue and an upper bound computed by training OC-SVM on 500 support samples, thus its results are the same for 1/5 shot. Top results are in bold.

### 3.3. Results

**FSOCC:** Table 1 summarizes the results for FSOCC on miniImagenet and tieredImageNet. There are two main observations. 1) Our HyperClass reaches SoTA performance with much shallower (vanilla) Conv-4 backbone (64 features), used also in (19) 2) Employing a deeper backbone, our simple baselines already surpass the previous methods while further boosting the performance on both datasets. Our transductive HyperClass variant allows additional performance gain, with a stronger impact on 1-shot. Finally, we add to the table an upper bound to this task, obtained by training a one-class SVM with 500 positive samples in the support set, instead of 1/5. The results are not far from our HC results, showing the strong capabilities of HC with few labeled samples.

**FSOR:** Table 2 shows the AUROC for negative detection (open-set recognition) task in FSOR. We reach a performance comparable to SoTA on miniImagenet and surpass previous work on tieredImageNet in 5-shot (with 82.83% vs. 81.64%) while being comparable to SoTA on 1-shot (with 74.54% vs. 74.95%). HyperClass also obtains lower STD values, indicating a higher statistical validity of our results.

Interestingly, our conducted simple baselines of OC-SVM and Proto show very competitive results, a fact that to the best of our knowledge has been overlooked so far.

| | miniImagenet 5-way | | tieredImageNet 5-way | |
|---|---|---|---|---|
| Model | 1-shot | 5-shot | 1-shot | 5-shot |
| PEELER (14) | 60.57±0.83 | 67.35 | 65.20 | 73.27 |
| SnaTCHer-F (5) | 68.27±0.96 | 77.42 | 74.28 | 82.02 |
| SnaTCHer-T (10) | 70.17±0.88 | 76.66 | 74.84 | <u>82.03</u> |
| SnaTCHer-L (10) | 69.40±0.92 | 76.15 | **74.95** | 80.81 |
| ATT (9) | <u>71.35±0.68</u> | <u>79.85±0.58</u> | 72.74±0.78 | 78.66±0.65 |
| ATT-G (9) | **72.41±0.72** | **79.85±0.57** | 73.43±0.78 | 81.64±0.63 |
| OC-SVM | 69.26±0.68 | 72.58±0.59 | 71.58±0.57 | 75.00±0.49 |
| Proto | 70.59±0.64 | 76.74±0.55 | 74.85±0.53 | 81.40±0.42 |
| HC (Ours) | 70.29±0.26 | 79.21±0.50 | <u>74.54±0.54</u> | **82.83±0.38** |

Table 2: AUROC for FSOR in-set classification subtask (%) with confidence interval of 95% over 600 episodes. All methods use ResNet12 backbone resulting in feature dimension 640D. Top performance are in bold, and 2nd place are underlined.

## 4. Summary

One class classification and open set recognition need to handle out-of-distribution data. In few-shot scenarios, the model must learn from just a handful (1-5) of samples to discriminate both uni-modal and multi-modal in-set categories from external instances. This paper suggests a new method for learning a one class and an open set task, that is further extended to transductive learning. Our evaluation shows SoTA results for few-shot one-class classification with competitive performance on open-set recognition task.

## References


[1] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, pages 1563–1572, 2016. 1

[2] Van Loi Cao, Miguel Nicolau, and James McDermott. One-class classification for anomaly detection with kernel density estimation and genetic programming. In *Genetic Programming: 19th European Conference, EuroGP 2016, Porto, Portugal, March 30-April 1, 2016, Proceedings 19*, pages 3–18. Springer, 2016. 1

[3] Gabriel Dahia and Maurício Pamplona Segundo. Meta learning for few-shot one-class classification. *AI*, 2(2):195–208, 2021. 1

[4] Jia Deng, Wei Dong, Richard Socher, LiJia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 3

[5] Ahmed Frikha, Denis Krompaß, Hans-Georg Köpken, and Volker Tresp. Few-shot one-class classification via meta-learning. In *AAAI*, volume 35, pages 7448–7456, 2021. 1, 3, 4

[6] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020. 1

[7] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004. 3

[8] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *CVPR*, pages 9068–9077, 2022. 3

[9] Shiyuan Huang, Jiawei Ma, Guangxing Han, and Shih-Fu Chang. Task-adaptive negative envision for few-shot open-set recognition. In *CVPR*, pages 7171–7180, 2022. 1, 3, 4

[10] Minki Jeong, Seokeon Choi, and Changick Kim. Few-shot open-set recognition by transformation consistency. In *CVPR*, pages 12566–12575, 2021. 1, 3, 4

[11] Jedrzej Kozerawski and Matthew Turk. Clear: Cumulative learning for one-shot one-class image recognition. In *CVPR*, pages 3446–3455, 2018. 1

[12] Jedrzej Kozerawski and Matthew Turk. One-class meta-learning: Towards generalizable few-shot open-set classification. *arXiv:2109.06859*, 2021. 1, 3, 4

[13] Anna Kruspe. One-way prototypical networks. *arXiv:1906.00820*, 2019. 1

[14] Bo Liu, Hao Kang, Haoxiang Li, Gang Hua, and Nuno Vasconcelos. Few-shot open-set recognition using meta-learning. In *CVPR*, pages 8798–8807, 2020. 1, 3, 4

[15] Janaina Mourão-Miranda, David R Hardoon, Tim Hahn, Andre F Marquand, Steve CR Williams, John Shawe-Taylor, and Michael Brammer. Patient classification as an outlier detection problem: an application of the one-class support vector machine. *Neuroimage*, 58(3):793–804, 2011. 1

[16] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICML*, 2017. 3

[17] 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. *ICML*, 2018. 3

[18] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 1

[19] 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 *ECCV*, pages 266–282, 2020. 2, 3, 4

[20] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 3