

Improving Few-Shot Learning using Composite Rotation based Auxiliary Task

Pratik Mazumder[†]Pravendra Singh[†]Vinay P. Namboodiri^{†*}[†]Department of Computer Science and Engineering, IIT Kanpur, India^{*}University of Bath, United Kingdom

{pratikm, psingh}@iitk.ac.in, vpn22@bath.ac.uk

Abstract

In this paper, we propose an approach to improve few-shot classification performance using a composite rotation based auxiliary task. Few-shot classification methods aim to produce neural networks that perform well for classes with a large number of training samples and classes with less number of training samples. They employ techniques to enable the network to produce highly discriminative features that are also very generic. Generally, the better the quality and generic-nature of the features produced by the network, the better is the performance of the network on few-shot learning. Our approach aims to train networks to produce such features by using a self-supervised auxiliary task. Our proposed composite rotation based auxiliary task performs rotation at two levels, i.e., rotation of patches inside the image (inner rotation) and rotation of the whole image (outer rotation) and assigns one out of 16 rotation classes to the modified image. We then simultaneously train for the composite rotation prediction task along with the original classification task, which forces the network to learn high-quality generic features that help improve the few-shot classification performance. We experimentally show that our approach performs better than existing few-shot learning methods on multiple benchmark datasets.

1. Introduction

Deep learning techniques have been used extensively to tackle several computer vision tasks, and they have been very successful [32, 18, 47]. The ability of neural networks to learn informative features from images is the main factor behind the successes of deep learning frameworks. However, neural networks need to be trained on large volumes of labeled data, which is a cause of concern since obtaining labeled data can be very difficult and sometimes very expensive. Obtaining labeled data may also require manual annotation, which is time-consuming and costly. In many real-world cases, it is not feasible to collect a large amount of labeled data for all categories of data. In such a case, the

network will not perform well for classes with few labeled training examples. On the other hand, humans can learn new categories of images from very few samples and recognize them in the wild with high probability. Deep learning models generally do not possess such a capability and are hence not human-like. Researchers have been looking into ways to achieve this. Few-shot learning is a step in this direction.

In the few-shot learning setting, the networks are trained in such a way that they can perform well for classes with few training examples and classes with many training examples. This can be achieved when the network has the ability to extract highly discriminative features from input images. The network should be trained in such a way that it can extract discriminative features even for a new set of categories. Few-shot learning methods generally operate on episodes. Episodes are tiny-datasets with a small train set and a small test set. Each episode consists of examples from a fixed small number of classes.

There have been many works in few-shot learning. Prototypical network [37] computes class prototypes and then uses the nearest neighbor-based classification to predict classes for query images. MAML [10] trains the network to quickly adapt to a new set of classes to perform classification on them. LEO [33] learns to generate weights for classifier using the support examples of the classes in the episode. RFS [41] proposes to improve the quality of representation produced by the network by using knowledge distillation.

Since the performance of few-shot learning methods heavily depends on the discriminative and generic nature of the features extracted by its network, researchers have been looking to improve networks and enable them to extract such features. One standard method of improving the representation/features produced by a network is to use self-supervised learning. Self-supervised learning techniques do not require labeled data to train. Self-supervised learning is a type of unsupervised learning which involves creating artificial labels using unlabeled examples. Artificial labels are created in a way that is simple, fast, and automated, us-

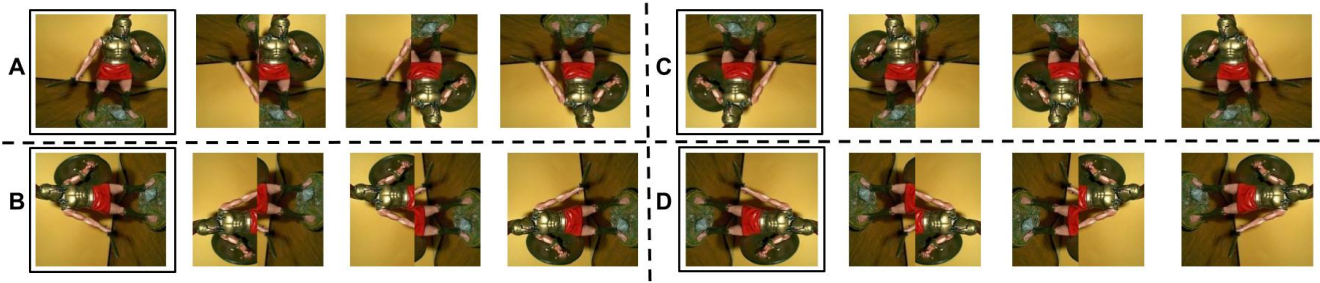


Figure 1. **Composite rotation classes created from a single image.** The four images, which are inside a separate box, represent the 4 outer rotations used by [14]. Our composite rotation first rotates the entire image by 0, 90, 180, or 270 degrees (A,B,C,D) (outer rotation), then splits the image vertically from the middle and rotates each half by 0 or 180 degrees (inner rotation). This creates 16 combinations, which become the 16 classes of our self-supervision method.

ing only the visual data contained in the input. Training on such pseudo-labeled examples provides a pseudo supervised learning environment that helps the network to gain the ability to extract good and generic features from images. Self-supervised learning trains networks by making use of the structural information contained in the input images. Some widely used self-supervision techniques include training the network to predict the angle of rotation applied to the input image [14], to predict the relative position of a patch of an image with respect to a given patch of the image [7, 26], to produce color images from gray-scale images [19, 46].

Since self-supervised learning improves the features produced by a network and does not require additional labeled data, it is a great candidate to be used to improve few-shot learning. We propose to train the few-shot network on an auxiliary self-supervised task based on composite rotation. Our proposed composite rotation divides the image into patches and rotates them inside the image in addition to rotating the whole image (see Fig. 1). The network is trained to predict the type of composite rotation that has been applied to the image along with predicting the actual label. This multi-task training forces the model to learn more detailed features of the objects contained inside the images.

In order to validate the efficacy of our auxiliary task, we incorporate it into the method (RFS) proposed in [41]. RFS makes use of knowledge distillation to improve the representational quality of the network, thereby improving its few-shot classification performance. We show experimentally that our auxiliary task can further significantly improve RFS. A simple rotation based auxiliary task is proposed in [12]. However, the simple rotation operation rotates the entire image as a whole (outer rotation). Therefore, the relative position between the points in the image does not change. This can be considered as a bias in the data. Due to this bias in the data, the features produced by the network trained using this technique can still be improved to capture more meaningful details about the objects in the image. This can be done by rotating patches inside the image (in-

ner rotation) in addition to the outer rotation. Our proposed composite rotation performs both inner and outer rotations. For a fair comparison with [12], we replace their rotation based auxiliary task with our composite rotation based auxiliary task (CRAT) and perform few-shot learning experiments. We experimentally show that our proposed auxiliary task performs significantly better than the simple rotation based auxiliary task used in [12]. Our approach is described in detail in Sec 3.

The contributions of this paper are as follows:

- We propose an approach to improve few-shot classification that trains the network on a composite rotation based auxiliary self-supervised task along with the general classification task.
- We incorporate our proposed auxiliary task into multiple few-shot learning methods [41, 12] and experimentally show that we are able to improve significantly over them.
- We perform experiments on multiple benchmark few-shot learning datasets to show the efficacy of our approach. We also perform ablation experiments to validate our approach.

2. Related Works

2.1. Few-Shot Learning

Many few-shot learning methods have been proposed by researchers [10, 37, 11, 13, 39, 3, 41, 36, 15, 21]. Prototypical Network [37] first averages the embeddings of the support data-points of each class produced by a base network to obtain a “prototype” for each class and then finds the class embedding closest to the embedding of the query image. Model Agnostic Meta Learning [10] optimizes the network in such a way that it can quickly adapt to a new episode. A graph neural network architecture is used in [34]. Meta Network [25] performs fast parameterization of the underlying network for rapid generalization.

The method in [31] applies the prototypical network to a semi-supervised setup by making use of labeled and unlabeled examples in each episode. TADAM [27] uses a task-based embedding as an attention to the convolutional layers of the base network. This attention helps shift the image embeddings closer to the embeddings of similar images based on the classes that are being trained on. RelationNet [39] uses relation scores to match query and support images of the classes. Learning without forgetting [13] uses support examples of the novel classes and classifier weights of the base classes to learn classifier weights for the novel classes. R2D2 [1] makes use of fast convergent methods like ridge regression for few-shot learning. LEO [33] solves an optimization problem in the parameter space to learn good parameters for the few-shot classifier.

MetaOptNet [20] learns more discriminative features by making use of linear predictors. TPN [23] uses a graph-based method to exploit the test data itself to improve few-shot classification in a transductive setting. In [36], the authors propose to use dynamic classifiers constructed from limited samples. The method proposed in [15] improves the generation of classifier weights for few-shot classification by maximizing the mutual information between the weights and the data. In [21], the authors use conditional Wasserstein GAN to hallucinate highly discriminative features.

RFS [41] makes use of self-distillation, which is a knowledge transfer process from a trained network to a student network with the same architecture. The authors show that this results in the student network learning better and more generic features. Such features enable the student network to be more generalizable and perform better on few-shot learning. Recently, self-supervision has been applied to few-shot learning. In [12], the authors use self-supervised techniques as auxiliary tasks to improve few-shot learning. The proposed method trains the model simultaneously on an auxiliary self-supervised task in addition to its original task. The images are modified to create self-supervised tasks such as rotation or relative patch position prediction. The authors show that using such auxiliary tasks helps the network perform better in the few-shot learning settings, indicating that such a network extracts a more generic set of features from the images.

Our paper focuses on training the network using a composite rotation based auxiliary task (CRAT) along with the classification task. Our composite rotation is more sophisticated than the simple rotation used in [12] and performs better than it for all the benchmark datasets that we experiment on. We also use our technique to improve RFS [41] and achieve state-of-the-art results.

2.2. Self-Supervised Learning Techniques

Many self-supervision techniques have been proposed to improve semantic feature learning. In [28], the authors train

the network to perform image inpainting or image completion. Several works have proposed to use the variance of image colorization as self-supervision [19, 46]. The method proposed in [7] trains the network to predict the relative position of a patch of an image with respect to a given patch of the same image. A convolutional neural network is trained to solve Jigsaw puzzles in [26].

In [14], the authors rotate images by a fixed set of angles and then train the network to predict the angle of rotation. The method proposed in [8] forms surrogate classes by transforming images in specific ways and trains convolutional neural networks to predict the class of the image. In [9], the authors propose to train the network to learn additional discriminative features along with the rotation angle prediction. This is achieved by additionally training the network to reduce the distance between different versions of the same image, which have been rotated at a different angle. This ensures that the network learns to produce features that have a better instance-level discriminative ability.

Contrastive Multiview Coding (CMC) [40] takes different views of an image and trains the network to learn a representation that maximizes the mutual information between the different views. But CMC requires specialized architecture, including separate encoders for different views of the data. Momentum Contrast (MoCo) [16] performs matching of encoded queries q to a dictionary of encoded keys using a contrastive loss in order to train the network. But MoCo requires a memory bank to store the dictionary. SimCLR [4] applies separate sets of data augmentation to the input resulting in two different but correlated views and uses contrastive loss to bring them closer in the feature space. It does not require specialized architectures or a memory bank and still achieves state-of-the-art unsupervised learning results, outperforming CMC and MoCo. We compare an auxiliary task based on SimCLR with our proposed composite rotation based auxiliary task in Sec. 4.7.2.

3. Method

3.1. Problem Setting

In the few-shot learning setting, the train and test classes are referred to as base and novel classes, respectively. Here, the networks operate on episodes of data. An episode can be thought of as a small dataset that is further divided into a mini-train set and a mini-test set. Each episode consists of a fixed small number of classes N , and each class has K support training examples. Such episodes are known as N -way K -shot episodes. The test set consists of query data points belonging to one of the N classes.

3.2. Proposed Composite Rotation based Auxiliary Task (CRAT)

We propose a composite rotation based auxiliary task (CRAT) that helps in improving few-shot learning perfor-

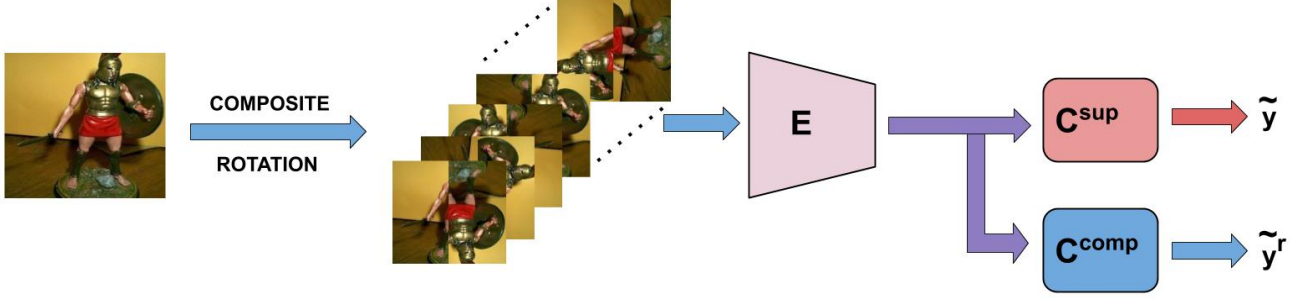


Figure 2. Self-supervised auxiliary classification task. The input image is rotated using our composite rotation technique. The rotated images are fed to the feature extraction unit E . C^{sup} and C^{comp} perform fully-supervised classification and auxiliary composite rotation classification on the features extracted by E .

mance. In our proposed composite rotation, we first rotate the entire image by either 0, 90, 180, or 270 degrees. Next, we split the image vertically from the middle. Then, we rotate each half by either 0 or 180 degrees (see Fig. 1). The outer rotation angle remains the same for a total of 4 combinations, while the inner rotation on each half is either 0 or 180 degrees. Therefore, there are a total of 16 classes of composite rotation. We validate the operations used in this composite rotation through ablation experiments in Sec 4.7.1. We use this composite rotation as an auxiliary task to improve few-shot classification methods. A few-shot learning network generally consists of a feature extraction network E , that extracts features from images. We add a composite rotation classification network C^{comp} for our proposed auxiliary task. For any given input image, we take the output of the feature extraction network E and feed it to C^{comp} to predict the composite rotation class. The training on this auxiliary task is carried out in parallel to the main classification task. After the training process is completed, the composite rotation classification network C^{comp} is discarded as it has served its purpose of improving the feature extraction network E . The composite rotation based auxiliary task loss L_{CR} can be defined as

$$L_{CR}(E, C^{comp}) = \sum_i f_{CE}(\tilde{y}_i^r, y_i^r) \quad (1)$$

where i refers to the i^{th} data point in the mini-batch, f_{CE} refers to the cross-entropy loss function, \tilde{y}_i^r refers to the composite rotation class of the i^{th} data point predicted by C^{comp} , and y_i^r refers to the actual composite rotation class of the i^{th} data point.

Our composite rotation technique will cause some objects within the image to behave differently after rotation since the rotation is not being applied only to the entire image as a whole. There will be cases where semantic parts of the same object in the image will get rotated differently. Objects may also end up rotating to different centers of rotation. Therefore, the relative position of objects in the image and even the relative position of the semantic parts of the same object may get changed after our composite rotation.

This will force the network to extract more detailed features about the objects, their parts, and their orientations. As a result, networks trained on these types of pseudo-classes should be better at extracting meaningful features from images.

Ideally, we should create even more subdivisions within the image and rotate them in order to complicate the task even more. However, this will exponentially increase the total number of classes and, hence, the model’s complexity, which might make it difficult to train. We tried to split the image into four quarters and rotate each of them through the four angles apart from the full image rotation, which led to a total of 1024 classes. The resulting auxiliary task is very difficult, and the network performance is not adequate on few-shot classification, as shown in Sec. 4.7.1.

3.3. Integrating CRAT with RFS [41]

In [41], the authors propose a few-shot learning method (RFS) focused on improving the feature representation quality of the network. The model consists of a feature extraction network E , a fully-supervised classification network C^{sup} , and a linear few-shot classification module C^{few} . The method proposed in [41] first trains the feature extraction network E and the fully-supervised classification network C^{sup} on the entire training set (phase 1). The trained network from phase 1 is used as a teacher (E_T, C_T^{sup}) to train a student network (E_S, C_S^{sup}) with the same architecture using self-distillation multiple times (phase 2). For self-distillation, the student is trained to minimize the Kullback-Leibler divergence (KL) between the soft predictions of the teacher and the student networks. During phase 2, the student is trained on the fully supervised classification loss on the entire training set and on the self-distillation loss. C_S^{sup} is discarded after the training process. During testing, for every testing episode, E_S is used to extract features for the support examples of each class. These features are used to train a logistic regression classification model C_S^{few} . Finally, E_S is used to extract the query sample features, and C_S^{few} is used to classify them.

We introduce our composite rotation based auxiliary task (CRAT) into this method to improve it. A composite rotation classification network C^{comp} is added to the model. During training, E is used to extract features for the rotated image. This feature is fed to both C^{sup} and C^{comp} to perform fully supervised classification and composite rotation classification, respectively (see Fig. 2). This will help the feature extractor learn to extract better features in order to perform better for both the tasks.

In this modified training procedure, the phase 1 now involves training the feature extraction network E , the fully-supervised classification network C^{sup} and the composite rotation classification network C^{comp} on the entire training set. In phase 2, the trained network from phase 1 is used as a teacher ($E_T, C_T^{sup}, C_T^{comp}$) to train a student network ($E_S, C_S^{sup}, C_S^{comp}$) with the same architecture using the fully-supervised classification loss, the composite rotation prediction loss, and the self-distillation loss. During testing, both C_S^{sup} and C_S^{comp} are discarded. E_S extracts features for the support examples of each class, which are used to train a logistic regression classification model C_S^{few} . Finally, C_S^{few} is used to classify the query samples. We experimentally show that our proposed composite rotation based auxiliary task can significantly improve the performance of RFS. The new phase 1 and 2 losses can be defined as:

$$L_{phase1} = \sum_i (f_{CE}(\tilde{y}_i, y_i) + \lambda f_{CE}(\tilde{y}_i^r, y_i^r)) \quad (2)$$

where, y_i is the real label for the i^{th} data point, \tilde{y}_i refers to the label predicted by C^{sup} for the i^{th} data point, \tilde{y}_i^r refers to the composite rotation class predicted by C^{comp} for the i^{th} data point, y_i^r refers to the actual composite rotation class of the i^{th} data point, λ hyper-parameter determines the influence of the auxiliary task loss in the total loss.

$$L_{phase2} = \sum_i [\alpha (f_{CE}(\tilde{y}_i, y_i) + \lambda f_{CE}(\tilde{y}_i^r, y_i^r)) + \beta KL(\phi_S(x_i), \phi_T(x_i))] \quad (3)$$

where, KL refers to the KL-Divergence function, α, β are hyper-parameters, $\phi_S(x_i), \phi_T(x_i)$ refer to the soft label predictions of the student and teacher models for the i^{th} data point. $\phi_S(x_i) = F_{softmax}(C_S^{sup}(E_S(x_i)))$, $\phi_T(x_i) = F_{softmax}(C_T^{sup}(E_T(x_i)))$ where $F_{softmax}$ is the softmax function.

3.4. Integrating CRAT with BF3S [12]

The method (BF3S) proposed in [12] also uses an auxiliary task. It uses a simple rotation [14] based self-supervised auxiliary task to train the feature extraction network simultaneously along with the few-shot classification task. After the training step, the auxiliary task network is discarded, and the remaining network is used to perform

few-shot classification. The simple rotation based self-supervised task rotates the image by a fixed angle (out of $0^\circ, 90^\circ, 180^\circ, 270^\circ$), and then trains the network to predict the angle of rotation. This does not change the relative position of points inside the image, and all points rotate with the same center of rotation, i.e., the center of the image. This leads to a bias in the data. We replace this simple rotation based auxiliary task with our composite rotation based auxiliary task and perform the same training process as proposed in BF3S. We experimentally show that our composite rotation based auxiliary task performs significantly better than the auxiliary tasks originally used in this method.

4. Experiments

4.1. Datasets

We perform few-shot classification experiments on 4 benchmark datasets: mini-ImageNet [42], tiered-ImageNet [31], CIFAR-FS [1] and FC-100 [27].

mini-ImageNet [42] consists of 100 classes, each of which has around 600 images of size 84×84 pixels. The classes are divided into 64 train classes, 16 validation classes, and 20 test classes. It has been derived from the ImageNet [32] dataset. tiered-ImageNet [31] is also derived from ImageNet and consists of 351 train, 97 validation, and 60 test classes. CIFAR-FS and Few-shot-CIFAR100 (FC-100) are both derived from CIFAR-100 [17] dataset. CIFAR-FS consists of 64 train, 16 validation, and 20 test classes with images of size 32×32 pixels. FC-100 consists of 60 train, 20 validation, and 20 test classes with images of size 32×32 pixels. FC-100 splits classes based on the super-classes in CIFAR-100 in order to minimize similarity between classes from different splits.

4.2. Implementation Details

In order to show the efficacy of our proposed auxiliary task, we experiment with RFS [41] and BF3S [12]. We modify RFS [41] to include our proposed composite rotation based auxiliary task (CRAT), and we replace the auxiliary task in BF3S [12] with CRAT. For the experiments involving RFS [41], we use the ResNet-12 architecture for the feature extraction network E . It consists of 4 residual blocks, each having 3 convolutional layers with kernel size 3×3 . A max-pooling layer of kernel size 2×2 is added after each of the initial 3 residual blocks, and a global average pooling layer is added after the last residual block. The composite rotation classifier is implemented as a convolutional neural network with 4 convolutional blocks. Each block consists of a convolutional layer of 640 convolutional filters of kernel size of 3×3 , a batch normalization layer, and a ReLU activation function. Each convolutional block has 640 output filters. An adaptive average pooling block is added after the last convolutional block, and it is followed

by a fully-connected layer with input size 640 and output size 16. The λ hyper-parameter, that decides the contribution of our auxiliary task loss to the total loss is taken as 1 for all experiments. We take $\alpha = 0.5$ and $\beta = 1$ for the RFS experiments. During training, we create multiple copies of the same image with different composite rotation transformations. To keep the process scalable, we apply only five transformations per image at a time while classifying for all the 16 transformations. The other settings used for these experiments are the same as given in [41].

For the experiments involving BF3S [12], we conduct experiments on the best performing model of [12] that uses the WideResNet-28-10 [44] architecture as the feature extractor and the Cosine Classifier (CC). It is a 28-layer Wide Residual Network [44] with width factor 10, which produces a feature map of size $10 \times 10 \times 640$ and a global average pooling converts this to a 640-dimensional feature vector. For the composite-rotation based auxiliary task classification network, we use a 4-residual-layer residual block followed by a global average pooling block (similar to [12]) and a fully connected classification layer with an output size of 16. For the BF3S experiments, the settings are the same as given in [12]. The classification performance is estimated using an average of the classification accuracy over the test images achieved by the network in each task/episode. We present the results for the 5-way 1-shot and 5-shot settings.

4.3. mini-ImageNet Results

The results for few-shot classification on the mini-ImageNet dataset are given in Table 1. The results indicate that using CRAT significantly improves both BF3S [12] and RFS [41]. In the case of BF3S, CC + CRAT outperforms CC + Rot [12] by absolute margins of 0.94% and 1.05% for the 1-shot 5-way and 5-shot 5-way settings, respectively. RFS + CRAT performs better than RFS by absolute margins of 3.62% and 1.61% for the 1-shot 5-way and 5-shot 5-way settings, respectively. RFS + CRAT*, which is trained on the combined training and validation set, performs better than RFS* (train+val) by absolute margins of 2.31% and 1.64% for the 1-shot 5-way and 5-shot 5-way settings respectively. RFS + CRAT performs better than existing state-of-the-art methods for both train and train+val settings.

4.4. tiered-ImageNet Results

Table 2 reports the results for few-shot classification on the tiered-ImageNet dataset. BF3S with our proposed auxiliary task (CC + CRAT) performs better than CC + Rot [12]. RFS + CRAT performs significantly better than RFS [41]. Our RFS + CRAT outperforms existing state-of-the-art methods for both train and train+val settings as shown in Table 2.

Table 1. Average 1/5-shot 5-way few-shot classification accuracy over test images from the novel classes of the mini-ImageNet dataset. * indicate methods that train on a union of train and validation (train+val) set.

Models	Backbone	1-shot	5-shot
MAML [10] (ICML'17)	Conv-4-64	48.70 ± 1.84%	63.11 ± 0.92%
Proto Net [37] (NIPS'17)	Conv-4-64	49.42 ± 0.78%	68.20 ± 0.66%
MetaNet[25](ICML'17)	ResNet-12	57.10 ± 0.70%	70.04 ± 0.63%
LwoF [13] (CVPR'18)	Conv-4-64	56.20 ± 0.86%	72.81 ± 0.62%
RelationNet [39] (CVPR'18)	Conv-4-64	50.44 ± 0.82%	65.32 ± 0.70%
GNN [34] (ICLR'18)	Conv-4-64	50.30%	66.40%
SNAIL [24] (ICLR'18)	ResNet-12	55.71 ± 0.99%	68.88 ± 0.92%
Qiao et al. [29]* (CVPR'18)	WRN-28-10	59.60 ± 0.41%	73.74 ± 0.19%
TADAM [27] (NIPS'18)	ResNet-12	58.50 ± 0.30%	76.70 ± 0.30%
TPN [23] (ICLR'19)	Conv-4-64	55.51 ± 0.86%	69.86 ± 0.65%
R2-D2 [1] (ICLR'19)	Conv-4-64	49.50 ± 0.20%	65.40 ± 0.20%
R2-D2 [1] (ICLR'19)	Conv-4-512	51.80 ± 0.20%	68.40 ± 0.20%
STANet [43] (AAAI'19)	ResNet-12	58.35 ± 0.57%	71.07 ± 0.39%
IdeMe-Net [5] (CVPR'19)	ResNet-18	59.14 ± 0.86%	74.63 ± 0.74%
Shot-Free [30] (ICCV'19)	ResNet-12	59.04%	77.64%
SalNet Intra[45](CVPR'19)	ResNet-101	62.22 ± 0.87%	77.95 ± 0.65%
LEO* [33] (ICLR'19)	WRN-28-10	61.76 ± 0.08%	77.59 ± 0.12%
BF3S CC+Rot[12](ICCV'19)	WRN-28-10	62.93 ± 0.45%	79.87 ± 0.33%
MetaOptNet[20](CVPR'19)	ResNet-12	62.64 ± 0.61%	78.63 ± 0.46%
MetaOptNet*[20](CVPR'19)	ResNet-12	64.09 ± 0.62%	80.00 ± 0.45%
RFS[41]	ResNet-12	64.82 ± 0.60%	82.14 ± 0.43%
RFS[41]*	ResNet-12	66.58 ± 0.65%	83.22 ± 0.39%
Warp-MAML [11](ICLR'20)	Conv-4-64	52.30 ± 0.80%	68.40 ± 0.60%
D-SVS[2](AAAI'20)	ResNet-12	60.16 ± 0.47%	77.25 ± 0.15%
Deep DTN [3] (AAAI'20)	ResNet-12	63.45 ± 0.86%	77.91 ± 0.62%
AFHN [21] (CVPR'20)	ResNet-18	62.38 ± 0.72%	78.16 ± 0.56%
AWGIM[15] (CVPR'20)	WRN-28-10	63.12 ± 0.08%	78.40 ± 0.11%
DSN-MR[36] (CVPR'20)	ResNet-12	64.60 ± 0.72%	79.51 ± 0.50%
DSN-MR[36]* (CVPR'20)	ResNet-12	67.09 ± 0.68%	81.65 ± 0.69%
BF3S CC[12]+CRAT (Ours)	WRN-28-10	63.87 ± 0.47%	80.92 ± 0.34%
RFS[41] + CRAT (Ours)	ResNet-12	68.44 ± 0.60%	83.75 ± 0.41%
RFS[41] + CRAT* (Ours)	ResNet-12	68.89 ± 0.61%	84.86 ± 0.45%

Table 2. Average 1/5-shot 5-way few-shot classification accuracy over test images from the novel classes of the tiered-ImageNet dataset. * indicate methods that train on a union of train and validation (train+val) set.

Models	Backbone	1-shot	5-shot
MAML [10] (ICML'17)	Conv-4-64	51.67 ± 1.81%	70.30 ± 0.08%
Proto Net [37] (NIPS'17)	Conv-4-64	53.31 ± 0.89%	72.69 ± 0.74%
RelationNet [39] (CVPR'18)	Conv-4-64	54.48 ± 0.93%	71.32 ± 0.78%
TPN [23] (ICLR'19)	Conv-4-64	59.91 ± 0.94%	73.30 ± 0.75%
LEO* [33] (ICLR'19)	WRN-28-10	66.33 ± 0.05%	81.44 ± 0.09%
MetaOptNet [20](CVPR'19)	ResNet-12	65.99 ± 0.72%	81.56 ± 0.53%
MetaOptNet* [20](CVPR'19)	ResNet-12	65.81 ± 0.74%	81.75 ± 0.53%
Shot-Free [30] (ICCV'19)	ResNet-12	66.87%	82.64%
BF3S CC+Rot [12](ICCV'19)	WRN-28-10	70.53 ± 0.51%	84.98 ± 0.36%
RFS[41]	ResNet-12	71.52 ± 0.69%	86.03 ± 0.49%
RFS[41]*	ResNet-12	72.98 ± 0.71%	87.46 ± 0.44%
Warp-MAML[11] (ICLR'20)	Conv-4-64	57.20 ± 0.90%	74.10 ± 0.70%
AWGIM[15] (CVPR'20)	WRN-28-10	67.69 ± 0.11%	82.82 ± 0.13%
DSN-MR[36] (CVPR'20)	ResNet-12	67.39 ± 0.82%	82.85 ± 0.56%
DSN-MR[36]* (CVPR'20)	ResNet-12	68.44 ± 0.77%	83.32 ± 0.66%
BF3S CC[12]+CRAT(Ours)	WRN-28-10	71.37 ± 0.45%	85.89 ± 0.36%
RFS[41] + CRAT (Ours)	ResNet-12	73.45 ± 0.83%	87.33 ± 0.47%
RFS[41] + CRAT* (Ours)	ResNet-12	74.63 ± 0.78%	88.67 ± 0.44%

4.5. CIFAR-FS Results

Few-shot classification performance on the CIFAR-FS dataset are shown in Table 3. The results indicate that CC + CRAT performs better than CC + Rot by absolute margins of 4.17% and 2.81% for the 1-shot 5-way and 5-shot 5-way

Table 3. Average 1/5-shot 5-way few-shot classification accuracy over test images from the novel classes of the CIFAR-FS dataset. * indicate methods that train on a union of train and validation (train+val) set. †: results from [1]. ‡: results from [12].

Models	Backbone	1-shot	5-shot
Proto Net [37]† (NIPS'17)	Conv-4-64	55.50 ± 0.70%	72.00 ± 0.60%
Proto Net [37]† (NIPS'17)	Conv-4-512	57.90 ± 0.80%	76.70 ± 0.60%
Proto Net [37]‡ (NIPS'17)	Conv-4-64	62.82 ± 0.32%	79.59 ± 0.24%
Proto Net [37]‡ (NIPS'17)	Conv-4-512	66.48 ± 0.32%	80.28 ± 0.23%
MAML [10]† (ICML'17)	Conv-4-64	58.90 ± 1.90%	71.50 ± 1.00%
MAML [10]† (ICML'17)	Conv-4-512	53.80 ± 1.80%	67.60 ± 1.00%
RelationNet [39]† (CVPR'18)	Conv-4-64	55.00 ± 1.00%	69.30 ± 0.80%
GNN [34]† (ICLR'18)	Conv-4-64	61.90%	75.30%
GNN [34]† (ICLR'18)	Conv-4-512	56.00%	72.50%
R2-D2 [1] (ICLR'19)	Conv-4-64	62.30 ± 0.20%	77.40 ± 0.20%
R2-D2 [1] (ICLR'19)	Conv-4-512	65.40 ± 0.20%	79.40 ± 0.20%
Shot-Free [30] (ICCV'19)	ResNet-12	69.15%	84.70%
MetaOptNet [20](CVPR'19)	ResNet-12	72.60 ± 0.70%	84.30 ± 0.50%
MetaOptNet* [20](CVPR'19)	ResNet-12	72.80 ± 0.70%	85.00 ± 0.50%
BF3S CC+Rot [12](ICCV'19)	WRN-28-10	73.62 ± 0.31%	86.05 ± 0.22%
RFS[41]	ResNet-12	73.89 ± 0.80%	86.93 ± 0.50%
RFS[41]*	ResNet-12	75.40 ± 0.80%	88.20 ± 0.50%
DSN-MR[36] (CVPR'20)	ResNet-12	75.60 ± 0.90%	86.20 ± 0.60%
DSN-MR[36]* (CVPR'20)	ResNet-12	78.00 ± 0.90%	87.30 ± 0.60%
BF3S CC[12]+CRAT(Ours)	WRN-28-10	77.79 ± 0.32%	88.86 ± 0.23%
RFS[41] + CRAT (Ours)	ResNet-12	77.18 ± 0.38%	89.36 ± 0.25%
RFS[41] + CRAT* (Ours)	ResNet-12	78.71 ± 0.35%	89.78 ± 0.26%

Table 4. Average 1/5-shot 5-way few-shot classification accuracy over test images from the novel classes of the FC-100 dataset. * indicate methods that train on a union of train and validation (train+val) set.

Models	Backbone	1-shot	5-shot
Proto Net [37] (NIPS'17)	Conv-4-64	35.30 ± 0.60%	48.60 ± 0.60%
TADAM [27] (NIPS'18)	ResNet-12	40.10 ± 0.40%	56.10 ± 0.40%
MetaOptNet [20](CVPR'19)	ResNet-12	41.10 ± 0.60%	55.50 ± 0.60%
MetaOptNet* [20](CVPR'19)	ResNet-12	47.20 ± 0.60%	62.50 ± 0.60%
MTL [38] (CVPR'19)	ResNet-12	45.10 ± 1.80%	57.60 ± 0.90%
DC [22] (CVPR'19)	ResNet-12	42.04 ± 0.17%	57.63 ± 0.23%
RFS[41]	ResNet-12	44.57 ± 0.70%	60.91 ± 0.60%
RFS[41]*	ResNet-12	51.60 ± 0.70%	68.40 ± 0.60%
Transductive[6] (ICLR'20)	WRN-28-10	43.16 ± 0.59%	57.57 ± 0.55%
Transductive[6]* (ICLR'20)	WRN-28-10	50.44 ± 0.68%	65.74 ± 0.60%
RFS[41] + CRAT (Ours)	ResNet-12	46.85 ± 0.35%	63.56 ± 0.33%
RFS[41] + CRAT* (Ours)	ResNet-12	54.70 ± 0.38%	70.76 ± 0.35%

settings, respectively. RFS + CRAT outperforms RFS by absolute margins of 3.29% and 2.43% for the 1-shot 5-way and 5-shot 5-way settings. RFS + CRAT*, with the model trained on train+val, performs better than RFS* (train+val) by absolute margins of 3.31% and 1.58% for the 1-shot 5-way and 5-shot 5-way settings respectively. The results also indicate that RFS + CRAT achieves state-of-the-art results.

4.6. FC-100 Results

Table 4 depicts the results for few-shot classification on the FC-100 dataset. From the results, it can be seen that RFS + CRAT performs better than RFS by absolute margins of 2.28% and 2.65% for the 1-shot 5-way and 5-shot 5-way settings, respectively. RFS + CRAT*, with the model trained on train+val, performs better than RFS* (train+val) by absolute margins of 3.1% and 2.36% for the 1-shot 5-way and 5-shot 5-way settings respectively. It also performs

Table 5. Average 1/5-shot 5-way few-shot classification accuracy over test images from the novel classes of the CIFAR-FS for auxiliary tasks based on different transformations. Experiments are conducted using RFS and ResNet-12 architecture.

Models	Classes	1-shot	5-shot
RFS + Rot	4	74.71 ± 0.69%	87.92 ± 0.26%
RFS + HS4	4	74.12 ± 0.66%	87.20 ± 0.27%
RFS + VS4	4	74.45 ± 0.65%	87.50 ± 0.25%
RFS + HS16	16	77.02 ± 0.35%	89.15 ± 0.26%
RFS + VS16 (CRAT)	16	77.18 ± 0.38%	89.36 ± 0.25%
RFS + Rot32 (HS16 + VS16)	32	77.31 ± 0.41%	89.51 ± 0.29%
RFS + Rot256	256	73.56 ± 0.40%	86.05 ± 0.28%
RFS + Rot1024	1024	73.77 ± 0.72%	86.36 ± 0.45%

better than existing state-of-the-art methods by a significant margin.

4.7. Ablations

We perform ablation experiments: 1) to validate the transformations involved in our composite rotation operation 2) to compare our composite rotation to other self-supervised auxiliary tasks.

4.7.1 Transformations in Composite Rotation

Our proposed composite rotation involves rotating the entire image by either 0, 90, 180, or 270 degrees, splitting the image vertically from the middle, and finally, rotating each half by either 0 or 180 degrees (see Fig. 1). We perform ablations to validate our transformation choices by using various combinations of transformations as the auxiliary task used along with RFS on the CIFAR-FS dataset with ResNet-12 architecture. RFS + Rot uses the simple rotation based self-supervision as used in [12] and has 4 classes. RFS + HS4 splits the image horizontally from the middle and rotates each half by either 0 or 180 degrees, resulting in 4 classes. Similarly, RFS + VS4 splits the image vertically from the middle and rotates each half by either 0 or 180 degrees. RFS + HS16 and RFS + VS16 rotate the entire image by either 0, 90, 180, or 270 degrees apart from splitting the image horizontally and vertically, respectively, and rotating each half by 0 or 180 degrees. This results in 16 classes. RFS + Rot32 combines RFS + HS16 and RFS + VS16 classes to obtain 32 classes of transformation. RFS + Rot256 splits the image into 4 quarters and rotates each quarter by either 0, 90, 180, or 270 degrees resulting in 256 classes. Similarly, RFS + Rot1024 splits the image into 4 quarters and rotates each quarter by either 0, 90, 180, or 270 degrees and also rotates the entire image by either 0, 90, 180, or 270 degrees resulting in 1024 classes.

Table 5 indicates that RFS + HS4 and RFS + VS4 perform close to RFS + Rot. Therefore, splitting an image into half and rotating each half performs similar to the simple rotation operation, when used as an auxiliary task. RFS + HS16 and RFS + VS16 perform significantly better than RFS + Rot. RFS + VS16 performs slightly better than RFS + HS16, and we choose VS16 for our composite rotation

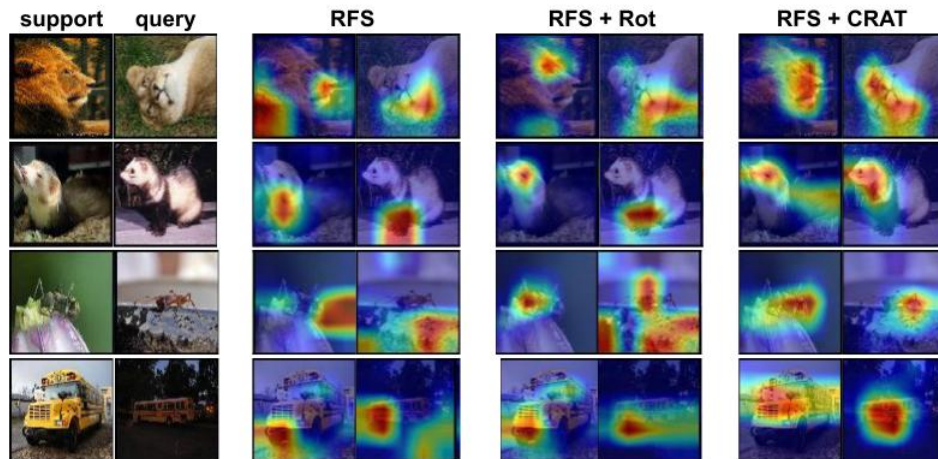


Figure 3. Class activation mapping visualization for novel class images of the mini-ImageNet dataset, using the feature extraction network of RFS, RFS + Rot (auxiliary rotation task), and RFS + CRAT (auxiliary composite rotation task) with ResNet-12 architecture.

Table 6. Average 1/5-shot 5-way few-shot classification accuracy over test images from the novel classes of the CIFAR-FS for different types of auxiliary self-supervised tasks on RFS with ResNet-12 architecture.

Models	1-shot	5-shot
RFS + Patch [7]	74.23 ± 0.65%	87.33 ± 0.35%
RFS + Rot [12, 14]	74.71 ± 0.69%	87.92 ± 0.26%
RFS + SimCLR [4]	74.45 ± 0.67%	87.42 ± 0.30%
RFS + CRAT (Ours)	77.18 ± 0.38%	89.36 ± 0.25%

operation. In RFS + Rot32, each image is converted to 32 different transformed samples. Since the performance of RFS + Rot32 is similar to RFS + VS16 and the memory overhead is high for RFS + Rot32, we use RFS + VS16 in our proposed approach. RFS + Rot256 and RFS + Rot1024 convert each image into 256 and 1024 different transformed samples respectively. This requires huge computing resources, which is not always feasible. In order to implement them in a scalable way, we randomly apply only 1 transformation class out of 256 in the case of Rot256 and 1 out of 1024 in the case of Rot1024 to each image. The results indicate that Rot1024 and Rot256 are unable to provide improvements to the base model. This possibly results from the large number of classes in the rotation classifier, which makes the auxiliary task very complicated and difficult and consequently hurts the classifier’s performance since they share the same feature extraction network.

4.7.2 Comparison with Other Self-Supervised Auxiliary Tasks

We compare our proposed composite rotation based auxiliary task (CRAT) with other auxiliary tasks based on self-supervised techniques. We perform experiments on RFS [41] with the auxiliary task as relative patch location prediction [7] (RFS + Patch), simple rotation (RFS + Rot) [12, 14] and SimCLR [4]. SimCLR is the current state-of-the-art self-supervised training technique that uses contrastive learning. SimCLR can be easily used as an auxiliary

task, which is a necessary pre-condition for our approach.

Table 6, shows that RFS + CRAT performs significantly better than RFS + Patch, RFS + Rot and RFS + SimCLR. Even though SimCLR is a state-of-the-art self-supervision method, it is unable to outperform CRAT when used as an auxiliary task for few-shot learning.

4.8. Qualitative Results

Fig. 3, shows the comparison of the class activation map [35] visualizations for RFS [41], RFS with simple rotation based auxiliary task (RFS + Rot) and RFS with our proposed composite rotation based auxiliary task (RFS + CRAT). The class activation mappings are visualized for the novel class images of the mini-ImageNet dataset, and the networks have not been trained on these classes. The visualizations show that RFS + CRAT attends more to the discriminative regions of the object, and this helps the network in extracting more discriminative and generic features, which in turn helps to improve the few-shot classification performance.

5. Conclusion

We propose a technique of improving few-shot classification by using a composite rotation based auxiliary task (CRAT). Our approach involves training the feature extraction network on this auxiliary task along with the general classification task. We demonstrate the efficiency of CRAT by plugging it into two recent few-shot learning methods [41, 12]. We experimentally show that our proposed auxiliary task significantly improves both of these few-shot learning methods. We show that RFS + CRAT outperforms existing state-of-the-art few-shot learning methods through several experiments on multiple benchmark datasets. We also validate our proposed composite rotation by performing ablation experiments.

References

- [1] Luca Bertinetto, Joao F. Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019.
- [2] Jiaxin Chen, Li-Ming Zhan, Xiao-Ming Wu, and Fu lai Chung. Variational metric scaling for metric-based meta-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
- [3] Mengting Chen, Yuxin Fang, Xinggang Wang, Heng Luo, Yifeng Geng, Xinyu Zhang, Chang Huang, Wenyu Liu, and Bo Wang. Diversity transfer network for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
- [4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.
- [5] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. Image deformation meta-networks for one-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8680–8689, 2019.
- [6] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *International Conference on Learning Representations*, 2020.
- [7] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.
- [8] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pages 766–774, 2014.
- [9] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning by rotation feature decoupling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10364–10374, 2019.
- [10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.
- [11] Sebastian Flennerhag, Andrei A. Rusu, Razvan Pascanu, Francesco Visin, Hujun Yin, and Raia Hadsell. Meta-learning with warped gradient descent. In *International Conference on Learning Representations*, 2020.
- [12] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8059–8068, 2019.
- [13] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018.
- [14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.
- [15] Yiluan Guo and Ngai-Man Cheung. Attentive weights generation for few shot learning via information maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13499–13508, 2020.
- [16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.
- [17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [19] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European Conference on Computer Vision*, pages 577–593. Springer, 2016.
- [20] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10657–10665, 2019.
- [21] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. Adversarial feature hallucination networks for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13470–13479, 2020.
- [22] Yann Lifchitz, Yannis Avrithis, Sylvaine Picard, and Andrei Bursuc. Dense classification and implanting for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9258–9267, 2019.
- [23] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sungju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *International Conference on Learning Representations*, 2019.
- [24] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *International Conference on Learning Representations*, 2018.
- [25] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2554–2563. JMLR. org, 2017.
- [26] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.
- [27] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved

- few-shot learning. In *Advances in Neural Information Processing Systems*, pages 721–731, 2018.
- [28] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.
- [29] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7229–7238, 2018.
- [30] Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Few-shot learning with embedded class models and shot-free meta training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 331–339, 2019.
- [31] Mengye Ren, Sachin Ravi, Eleni Triantafillou, Jake Snell, Kevin Swersky, Josh B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *International Conference on Learning Representations*, 2018.
- [32] 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*, 115(3):211–252, 2015.
- [33] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019.
- [34] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 2018.
- [35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [36] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. Adaptive subspaces for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4136–4145, 2020.
- [37] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.
- [38] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 403–412, 2019.
- [39] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.
- [40] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.
- [41] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020.
- [42] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.
- [43] Shipeng Yan, Songyang Zhang, Xuming He, et al. A dual attention network with semantic embedding for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9079–9086, 2019.
- [44] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.
- [45] Hongguang Zhang, Jing Zhang, and Piotr Koniusz. Few-shot learning via saliency-guided hallucination of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2770–2779, 2019.
- [46] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.
- [47] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.