Associative Alignment for Few-shot Image Classification

Arman Afrasiyabi*, Jean-François Lalonde*, and Christian Gagné*†
*Université Laval †Canada CIFAR AI Chair, Mila
arman.afrasiyabi.1@ulaval.ca
{jflalonde,christian.gagne}@gel.ulaval.ca

Abstract

Few-shot image classification aims at training a model by using only a few (e.g., 5 or even 1) examples of "novel" classes. The established way of doing so is to rely on a larger set of "base" data for either pre-training a model, or for training in a meta-learning context. Unfortunately, these approaches often suffer from overfitting since the models can easily memorize all of the novel samples. This paper mitigates this issue and proposes to leverage part of the base data by aligning the novel training instances to the closely related ones in the base training set. This expands the size of the effective novel training set by adding extra "related base" instances to the few novel ones, thereby allowing to train the entire network. Doing so limits overfitting and simultaneously strengthens the generalization capabilities of the network. We propose two associative alignment strategies: 1) a conditional adversarial alignment loss based on the Wasserstein distance; and 2) a metric-learning loss for minimizing the distance between related base samples and the centroid of novel instances in the feature space. Experiments on two standard datasets demonstrate that combining our centroid-based alignment loss results in absolute accuracy improvements of 4.4%, 1.2%, and 6.0% in 5-shot learning over the state of the art for object recognition, fine-grained classification, and cross-domain adaptation, respectively.

1. Introduction

Despite recent progress, generalizing on new concepts with little supervision is still a challenge in computer vision. In the context of image classification, few-shot learning aims to obtain a model that can learn to recognize novel image classes when very few training examples are available.

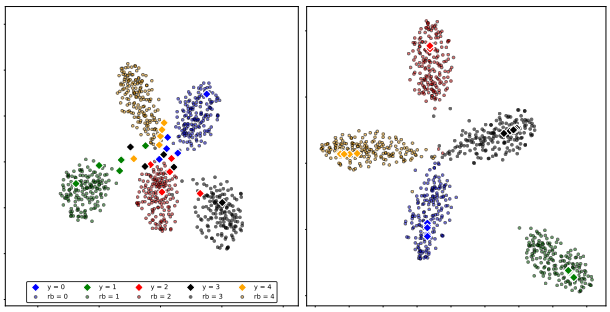
Meta-learning [7, 25, 29, 33] is typically used to tackle this problem, by extracting common knowledge from a large amount of labeled data (named the "base" classes) to train a model that can then learn to classify images from "novel" concepts with only a few examples. This is achieved by repeatedly sampling small subsets from the large pool of base images, effectively simulating the few-shot scenario.

Standard transfer learning has also been explored as an alternative solution [3, 11, 24]. The idea is to pre-train a network on the base samples and then fine-tune the classification layer on the novel examples. Interestingly, Chen *et al.* [3] demonstrated that doing so performs on par with the more sophisticated meta-learning strategies. It is, however, necessary to freeze the feature encoder part of the network when fine-tuning on the novel classes since the network otherwise overfits the novel examples. We hypothesize that this hinders performance and that gains could be made if the entire network is adapted to the novel, unseen categories.

In this paper, we propose an approach that simultaneously prevents overfitting without restricting the learning capabilities of the network for few-shot image classification. Our approach relies on the standard transfer learning strategy [3] as a starting point, but subsequently exploits base categories that are most similar (in the feature space) to the few novel samples to effectively provide additional training examples. We dub these similar categories the "related base" classes.

Of course, the related base classes represent different concepts than the novel classes, so fine-tuning directly on them could confuse the network (see fig. 1-(a)). A key contribution of our paper is thus to propose the *alignment* in the feature space of the novel examples with the related base samples (fig. 1-(b)). To this end, we propose and experimentally analyze two possible solutions for alignment: by 1) using a conditional parameterized critic network trained with the Wasserstein distance loss, or by 2) minimizing the distance between the related base samples to the centroid of the novel examples in the feature space. We demonstrate that by making such alignment, we can use related base classes along with the novel samples and obtain a significant improvement in few-shot classification performance.

We present the following contributions. First, we propose two approaches for aligning novel to related base classes in the feature space, allowing for effective training of entire networks for few-shot image classification. Second, we introduce a strong baseline that combines standard transfer learning [3] with an additive angular margin loss [6], along with early stopping to regularize the network while pre-training on the base categories. We find that this sim-



(a) before alignment (b) after alignment

Figure 1: The use of many related base examples (circles) in addition to few novel classes samples (diamonds) allows learning better discriminative models: (a) using directly similar examples from the base domain may not properly capture the novel classes; while (b) aligning both related base and novel training instances (in the feature space) provides more relevant training data for classification. Plots are generated with t-SNE [20] applied to the ResNet-18 feature embedding before (a) and after (b) the application of the centroid alignment algorithm. Points are color-coded by class.

ple baseline actually improves on the state of the art (both meta-learning and transfer learning), in the best case by 3% in overall accuracy. Third, we demonstrate through extensive experiments that our proposed centroid-based alignment significantly outperforms the state of the art in three types of scenarios: generic object recognition (gain of 4.4% in overall accuracy for 5-shot on *mini*-ImageNet), fine-grained classification (1.2% on CUB), and cross-domain adaptation (6.0% from *mini*-ImageNet to CUB).

2. Related work

The main few-shot learning approaches can be broadly categorized into meta-learning and standard transfer learning. In addition, data augmentation techniques (typically in meta-learning) have also been used for few-shot learning. We briefly review relevant works in each category below.

Meta-learning This family of approaches frames few-shot learning in the form of episodic training [7, 25, 29, 32]. An episode is defined by pretending to be in a few-shot regime while training on the base categories, which are available in large quantities. Initialization- and metric-based approaches are two variations on the episodic training scheme relevant for this work.

Initialization-based methods [7, 8, 17] learn parameters using base categories such that a small number of gradient steps over few examples. In contrast, our approach performs a larger number of updates, but require to maintain the alignment between the novel and their related base examples.

Metric-based approaches learn a metric with the intent of reducing the intra-class variations while training on base categories [2, 10, 16, 18, 19, 23, 29, 31, 33, 35, 37]. For example, Prototypical Networks [29] were proposed to learn a feature space where instances of a given class are located close to the corresponding prototype (centroid), allowing accurate distance-based classification. Our centroid alignment strategy borrows from the distance-based criteria of [29] but employs it for distribution matching in feature space rather than for building a classifier.

Standard transfer learning The strategy behind this method is to pre-train a network on the base classes and subsequently fine-tune it on the novel examples [3, 11, 24]. Despite its simplicity, Chen *et al.* [3] recently demonstrated that such an approach could result in similar generalization performance compared to meta-learning when deep backbones are employed as feature extractors. However, they have also shown that the weights of the pre-trained feature extractor must remain frozen while fine-tuning due to the propensity for overfitting. Although the training procedure we are proposing is similar to standard fine-tuning in base categories, our approach allows the training of the entire network, thereby increasing the learned model capacity while improving classification accuracy.

Data augmentation Another family of techniques relies on additional data for training in a few-shot regime, most of the time following a meta-learning training procedure [4, 5, 9, 12, 13, 21, 28, 36, 38]. Several ways of doing so have been proposed, including Feature Hallucination (FH) [13], which learns mappings between examples with an auxiliary generator that then hallucinates extra training examples (in the feature space). Subsequently, Wang *et al.* [36] proposed to use a GAN for the same purpose, and thus address the poor generalization of the FH framework. Unfortunately, it has been shown that this approach suffers from mode collapse [9]. Instead of generating artificial data for augmentation, Ren *et al.* [26] and Sun *et al.* [30] proposed semi-supervised meta-learning methods that make use of additional unlabeled data. We also rely on more data for training, but in contrast to these approaches, our method does not need any new data, nor does it require to generate any. Instead, we exploit the data that is *already available* in the base domain and get aligned the novel domain to the relevant base samples through fine-tuning.

3. Problem definition

We first present an overview of the transfer learning framework applied to few-shot classification, followed by the proposed approach.

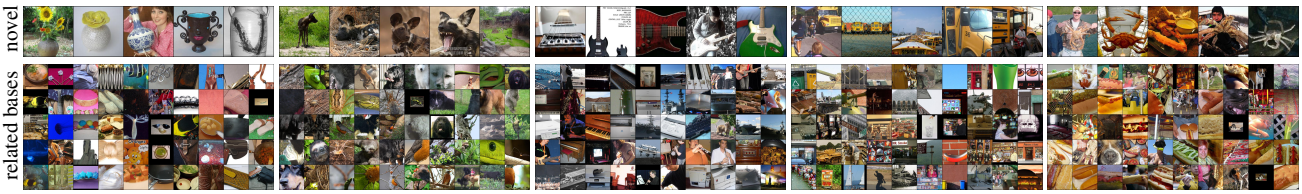


Figure 2: Results of related base algorithm in a 5-way 5-shot scenario. Each column represents a different novel class. The top row shows the 5 novel instances, while the bottom row show 60 randomly selected related base instances with $B = 10$.

3.1. Preliminaries

Let us assume that we have a large base dataset $\mathcal{X}^b = \{(\mathbf{x}_i^b, y_i^b)\}_{i=1}^{N^b}$, where $\mathbf{x}_i^b \in \mathbb{R}^d$ is the i-th data instance of the set and $y_i^b \in \mathcal{Y}^b$ is the corresponding class label. We are also given a small amount of novel class data $\mathcal{X}^n = \{(\mathbf{x}_i^n, y_i^n)\}_{i=1}^{N^n}$, with labels $y_i^n \in \mathcal{Y}^n$ from a set of distinct classes $\mathcal{Y}^n$. The goal of few-shot classification is to train a classifier with only a few examples from each of the novel classes (e.g., 5 or even just 1). In this work, we used the standard transfer learning strategy of Chen *et al.* [3], which is organized into the following two stages.

Pre-training stage The learning model is a neural network composed of a feature extractor $f(\cdot|\theta)$, parameterized by θ, followed by a linear classifier $c(\mathbf{x}|\mathbf{W}) \equiv \mathbf{W}^\top f(\mathbf{x}|\theta)$, described by matrix $\mathbf{W}$, ending with a scoring function such as softmax to produce the output. The network is trained from scratch on examples from the base categories $\mathcal{X}^b$.

Fine-tuning stage In order to adapt the network to the novel classes, the network is subsequently fine-tuned on the few examples from $\mathcal{X}^n$. Since overfitting is likely to occur if all the network weights are updated, the feature extractor weights θ are frozen, with only the classifier weights $\mathbf{W}$ being updated in this stage.

3.2. Our approach: associative alignment

While freezing the encoder weights θ during fine-tuning limits overfitting, it may also reduce the learning capacity of the model. We propose to maintain the original learning capacity by using a subset of *related* base categories, $\mathcal{X}^{rb} \subset \mathcal{X}^b$. To account for the discrepancy between the novel and related base classes, we propose to *align* the novel categories to the related base categories in the feature space. Such a mapping allows for a bigger pool of training data while making instances of these two sets more coherent.

Let us assume the set of instances $\mathcal{X}_i^n$ belonging to the i-th novel class $i \in \mathcal{Y}^n$, $\mathcal{X}_i^n = \{(\mathbf{x}_j^n, y_j^n) \in \mathcal{X}^n \,|\, y_j^n = i\}$, and the set of related base examples $\mathcal{X}_i^{rb}$ belonging to the same novel class i according to the $g(\cdot|\mathbf{M})$ mapping function, $\mathcal{X}_i^{rb} = \{(\mathbf{x}_j^b, y_j^b) \in \mathcal{X}^{rb} \,|\, g(y_j|\mathbf{M}) = i\}$. The function $g(y_j|\mathbf{M}) : \mathcal{Y}^b \to \mathcal{Y}^n$ maps base class labels to the novel ones according to the similarity matrix $\mathbf{M}$ (presented in the following subsection). We wish to find an alignment transformation for matching probability densities $p(f(\mathbf{x}_{i,k}^n \,|\, \theta))$ and $p(f(\mathbf{x}_{i,l}^{rb} \,|\, \theta))$. Here, $\mathbf{x}_{i,k}^n$ is the k-th element from class i in the novel set, and $\mathbf{x}_{i,l}^{rb}$ is the l-th element from class i in the related base set. This approach has the added benefit of allowing the fine-tuning of all of the model parameters θ and $\mathbf{W}$ with a reduced level of overfitting.

In the following of this section, we describe how the mapping between a novel class to a related base class (function $g(\cdot|\mathbf{M})$) can be found. Then, we propose two different alignment strategies using ideas borrowed from recent literature.

3.3. Detecting the related bases

We develop a simple, yet effective procedure to select a set of base categories related to a novel category. Our method associates B base categories to each novel class. After training $c(f(\cdot|\theta)|\mathbf{W})$ on $\mathcal{X}^b$, we first fine-tune $c(\cdot|\mathbf{W})$ on $\mathcal{X}^n$ while keeping θ fixed. Then, we define $\mathbf{M} \in \mathbb{R}^{K^b \times K^n}$ as a base-novel similarity matrix, where K^b and K^n are respectively the number of classes in $\mathcal{X}^b$ and $\mathcal{X}^n$. An element $m_{i,j}$ of the matrix $\mathbf{M}$ corresponds to the ratio of examples associated to the i-th base class that are classified as the j-th novel class:

$$m_{i,j} = \frac{1}{|\mathcal{X}_i^b|} \sum_{(\mathbf{x}_l^b, \cdot) \in \mathcal{X}_i^b} \mathbb{I}\left[j = \arg\max_{k=1}^{K^n} \left(c_k(f(\mathbf{x}_l^b|\theta) \,|\, \mathbf{W}) \right) \right], \tag{1}$$

where $c_k(f(\mathbf{x}|\theta)|\mathbf{W})$ is the classifier output $c(\cdot|\mathbf{W})$ for class k. From this base-novel similarity matrix $\mathbf{M}$, the related base to novel function, named $g(\cdot|\mathbf{M})$, is defined as:

$$\begin{aligned} &\mathcal{B}_0 = \mathcal{Y}^n, \qquad \mathcal{B}_i = B_{i-1} \backslash \mathcal{R}_i, \quad i = 1, \ldots, N^n, \\ &\mathcal{R}_i = \{j \in \mathcal{B}_{i-1} \,|\, \mathrm{rank}(j, \mathbf{M}_{[i,\mathcal{B}_{i-1}]}) \leq B\}, \\ &g(j|\mathbf{M}) = i \,|\, j \in \mathcal{R}_i. \end{aligned} \tag{2}$$

Function $\mathrm{rank}(j, \mathbf{m})$ provides the rank (integer) of value j in vector $\mathbf{m}$: a value of 1 means j is the largest value of $\mathbf{m}$. The B base classes with the highest score for a given novel class are kept as the related base for that class. We used a

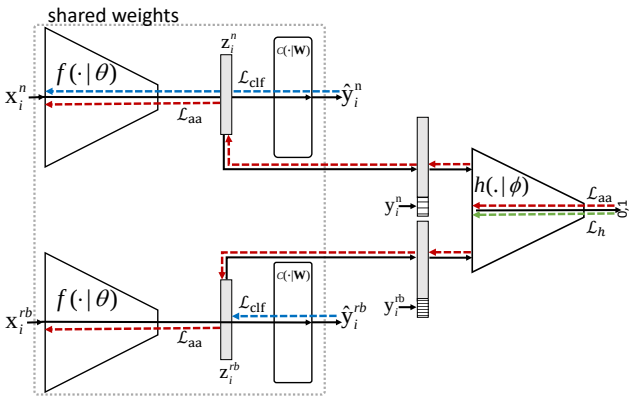


Figure 3: Overview of our adversarial alignment strategy. The adversarial alignment using feature learner $f(\cdot|\theta)$ takes an image $\mathbf{x}_i^n$ from i-th novel class and an example $\mathbf{x}_i^{rb}$ of related base from same class. The critic $h(\cdot|\phi)$ network takes the feature vectors and the corresponding one-hot vector of class label. Green arrow, red and blue arrows present the backpropagation of critic $\mathcal{L}_h$, adversarial $\mathcal{L}_{aa}$ and classification losses $\mathcal{L}_{\text{clf}}$ respectively. The dashed box represent a duplication of the networks that keep weights shared.

fixed number of associated categories B to ensure balance across the novel categories. Fig. 2 illustrates example results obtained with this method in a 5-shot, 5-way scenario.

4. Associative alignment strategies

We propose two associative strategies: adversarial and centroid alignment. We present each strategy in this section, and proceed to compare them experimentally in sec. 6.

4.1. Adversarial alignment

Inspired by WGAN [1], we propose to train the encoder $f(\cdot|\theta)$ to perform adversarial alignment using a conditioned critic network $h(\cdot|\phi)$ based on Wasserstein-1 distance between two probability densities p_x and p_y:

$$D(p_x, p_y) = \sup_{\|h\|_L \leq 1} \mathbb{E}_{x\sim p_y}[h(x)] - \mathbb{E}_{x\sim p_y}[h(x)]\,, \quad (3)$$

where sup is the supremum, and h is a 1-Lipschitz function.

Similarly to Arjovsky *et al.* [1], we use a parameterized critic network $h(\cdot|\phi)$ conditioned by the concatenation of the feature embedding of either $\mathbf{x}_i^n$ or $\mathbf{x}_j^{rb}$, along with the corresponding label y_i^n encoded as a one-hot vector. Conditioning $h(\cdot|\phi)$ helps the critic in matching novel categories and their corresponding related base categories. The critic

Algorithm 1: Adversarial alignment algorithm.

Input: pre-trained model $f(\cdot|\theta)$, classifier $c(\cdot|\mathbf{W})$, novel class set $\mathcal{X}^n$, related base set $\mathcal{X}^{rb}$
Output: aligned network $c(f(\cdot|\theta)|\mathbf{W})$
while *not done* **do**
  $\widetilde{\mathcal{X}}^n \leftarrow$ sample a batch from $\mathcal{X}^n$
  $\widetilde{\mathcal{X}}^{rb} \leftarrow$ sample a batch from $\mathcal{X}^{rb}$
  for i = 0,...,n_{critic} **do**
    evaluate critic loss $\mathcal{L}_h(\widetilde{\mathcal{X}}^n, \widetilde{\mathcal{X}}^{rb})$ with eq. 4
    update critic: $\phi \leftarrow \phi + \eta_h \nabla_\phi \mathcal{L}_h(\widetilde{\mathcal{X}}^n, \widetilde{\mathcal{X}}^{rb})$
    $\phi \leftarrow \text{clip}(\phi, -0.01, 0.01)$
  end
  evaluate alignment loss $\mathcal{L}_{\text{aa}}(\widetilde{\mathcal{X}}^n)$ with eq. 5
  $\theta \leftarrow \theta - \eta_{\text{aa}} \nabla_\theta \mathcal{L}_{\text{aa}}(\widetilde{\mathcal{X}}^n)$
  evaluate classification loss $\mathcal{L}_{\text{clf}}(\widetilde{\mathcal{X}}^{rb})$
  $\mathbf{W} \leftarrow \mathbf{W} - \eta_{\text{clf}} \nabla_{\mathbf{W}} \mathcal{L}_{\text{clf}}(\widetilde{\mathcal{X}}^{rb})$
  evaluate classification loss $\mathcal{L}_{\text{clf}}(\widetilde{\mathcal{X}}^n)$
  $\mathbf{W} \leftarrow \mathbf{W} - \eta_{\text{clf}} \nabla_{\mathbf{W}} \mathcal{L}_{\text{clf}}(\widetilde{\mathcal{X}}^n)$
  $\theta \leftarrow \theta - \eta_{\text{clf}} \nabla_\theta \mathcal{L}_{\text{clf}}(\widetilde{\mathcal{X}}^n)$
end

$h(\cdot|\phi)$ is trained with loss

$$\mathcal{L}_h(\mathcal{X}^n, \mathcal{X}^{rb}) = \frac{1}{N^{rb}} \sum_{(\mathbf{x}_i^{rb}, y_i^{rb}) \in \mathcal{X}^{rb}} h\left([f(\mathbf{x}_i^{rb}|\theta)\ y_i^{rb}]\,|\,\phi\right) - \frac{1}{N^n} \sum_{(\mathbf{x}_i^n, y_i^n) \in \mathcal{X}^n} h\left([f(\mathbf{x}_i^n|\theta)\ y_i^n]\,|\,\phi\right)\,, \quad (4)$$

where N^n and N^{rb} are the number of examples in $\mathcal{X}^n$ and $\mathcal{X}^{rb}$ respectively, and $[\cdot]$ is the concatenation operator, while the encoder parameters θ are updated using

$$\mathcal{L}_{\text{aa}}(\mathcal{X}^n) = \frac{1}{K^n} \sum_{(\mathbf{x}_i^n, y_i^n) \in \mathcal{X}^n} h\left([f(\mathbf{x}_i^n|\theta)\ y_i^n]|\phi\right)\,. \quad (5)$$

Algorithm 1 summarizes the training process of our adversarial alignment method. First, we perform the parameter update of critic $h(\cdot|\phi)$ using eq. 4. Similar to WGAN [1], we perform n_{critic} iterations to optimize h, before updating $f(\cdot|\theta)$ using eq. 5. Finally, the entire network is updated by a classification loss $\mathcal{L}_{\text{clf}}$ (defined in sec. 7).

4.2. Centroid alignment

We also propose a metric-based centroid distribution alignment strategy. The idea is to enforce intra-class compactness during the alignment process. Specifically, we explicitly push the training examples from the i-th novel

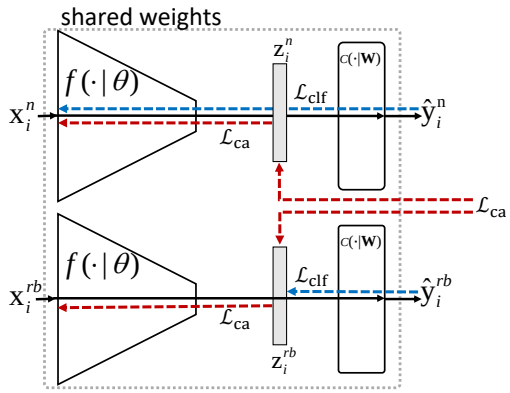


Figure 4: Schematic overview of our centroid alignment few-shot model. The feature learner $f(\cdot|\theta)$ takes an example from novel category $\mathbf{x}^n$ and an example related base $\mathbf{x}_i^{rb}$. A Euclidean centroid based alignment loss $\mathcal{L}_{\text{ca}}$ aligns the encoded $\mathbf{x}_i^n$ and $\mathbf{x}_i^{rb}$. Red arrows and blue arrows show the error backpropagation of centroid alignment loss $\mathcal{L}_{\text{ca}}$ and classification loss $\mathcal{L}_{\text{clf}}$ respectively.

class $\mathcal{X}_i^n$ towards the centroid of their related examples $\mathcal{X}_i^{rb}$ in feature space. The centroid $\boldsymbol{\mu}_i$ of $\mathcal{X}_i^{rb}$ is computed by

$$\boldsymbol{\mu}_i = \frac{1}{|\mathcal{X}_i^{rb}|} \sum_{(\mathbf{x}_j, y_h) \in \mathcal{X}_i^{rb}} f(\mathbf{x}_j|\theta),$$

which allows to define the centroid alignment loss as

$$\mathcal{L}_{\text{ca}}(\mathcal{X}^n) = \tag{6}$$
$$-\frac{1}{N^n N^{rb}} \sum_{i=1}^{K^n} \sum_{\mathbf{x}_j \in \mathcal{X}_i^n} \log \frac{\exp[-\|f(\mathbf{x}_j|\theta) - \boldsymbol{\mu}_i\|_2^2]}{\sum_{k=1}^{K^n} \exp[-\|f(\mathbf{x}_j|\theta) - \boldsymbol{\mu}_k\|_2^2]}.$$

Our alignment strategy bears similarities to [29] which also uses eq. 6 in a meta-learning framework. In our case, we use that same equation for matching distributions.

Fig. 4 illustrates our proposed centroid alignment, and algorithm 2 presents the overall procedure. First, we update the parameters of the feature extraction network $f(\cdot|\theta)$ using eq. 6. Second, the entire network is updated using a classification loss $\mathcal{L}_{\text{clf}}$ (defined in sec. 7).

5. Establishing a strong baseline

Before evaluating our alignment strategies in sec. 6, we first establish a strong baseline for comparison by following the recent literature. In particular, we build on the work of Chen *et al.* [3] but incorporate episodic early stopping on the pre-training stage with a different loss function.

5.1. Episodic early stopping

A fixed number of epochs in the pre-training stage has been commonly used (e.g., [3, 7, 29, 33]), but this might hamper performance when fine-tuning on novel classes. This is illustrated in fig. 5, which plots validation error after fine-tuning vs. the number of pre-training epochs. The "cosmax" function is used, with the entire network pre-trained on $\mathcal{X}^b$, and only the classification weights $\mathbf{W}$ fine-tuned on $\mathcal{X}^n$, as in [3]. The decrease in accuracy over the epochs (after 150 epoch for 1-shot) shows that pre-training should not be conducted for a fixed number of epochs. We thus propose to make use of episodic early stopping on the validation categories at pre-training time, more specifically by stopping the training when the mean accuracy on the validation over a window of recent epochs starts to decrease. The best model in the window is selected as the final result.

Algorithm 2: Centroid alignment algorithm.

Input: pre-trained model $f(\cdot|\theta)$, classifier $c(\cdot|\mathbf{W})$, novel class set $\mathcal{X}^n$, related base set $\mathcal{X}^{rb}$
Output: aligned network $c(f(\cdot|\theta)|\mathbf{W})$
while *not done* **do**
  $\widetilde{\mathcal{X}}^n \leftarrow$ sample a batch from $\mathcal{X}^n$
  $\widetilde{\mathcal{X}}^{rb} \leftarrow$ sample a batch from $\mathcal{X}^{rb}$
  evaluate alignment loss $\mathcal{L}_{\text{ca}}(\widetilde{\mathcal{X}}^n, \widetilde{\mathcal{X}}^{rb})$ with eq. 6
  $\theta \leftarrow \theta - \eta_{\text{ca}} \nabla_\theta \mathcal{L}_{\text{ca}}(\widetilde{\mathcal{X}}^n, \widetilde{\mathcal{X}}^{rb})$
  evaluate classification loss $\mathcal{L}_{\text{clf}}(\widetilde{\mathcal{X}}^{rb})$
  $\mathbf{W} \leftarrow \mathbf{W} - \eta_{\text{clf}} \nabla_{\mathbf{W}} \mathcal{L}_{\text{clf}}(\widetilde{\mathcal{X}}^{rb})$
  evaluate classification loss $\mathcal{L}_{\text{clf}}(\widetilde{\mathcal{X}}^n)$
  $\mathbf{W} \leftarrow \mathbf{W} - \eta_{\text{clf}} \nabla_{\mathbf{W}} \mathcal{L}_{\text{clf}}(\widetilde{\mathcal{X}}^n)$
  $\theta \leftarrow \theta - \eta_{\text{clf}} \nabla_\theta \mathcal{L}_{\text{clf}}(\widetilde{\mathcal{X}}^n)$
end

5.2. Classification loss functions

Deng *et al.* [6] have shown that an additive angular margin ("arcmax" hereafter) outperforms other metric learning algorithms for face recognition. The arcmax function has a metric learning property since it enforces a geodesic distance margin penalty on the normalized hypersphere, which we think can be beneficial for few-shot classification by helping keep classes cluster compact and well-separated.

Let $\mathbf{z} = f(\mathbf{x}|\theta)$ be the representation of $\mathbf{x}$ in feature space. As per [6], we transform the logit as $\mathbf{w}_j^\top \mathbf{z} = \|\mathbf{w}_j\| \|\mathbf{z}\| \cos \varphi_j$, where φ_j is the angle between $\mathbf{z}$ and $\mathbf{w}_j$, the j-th column in the weight matrix $\mathbf{W}$. Each weight $\|\mathbf{w}_j\| = 1$ by l_2 normalization. Arcmax adds an angular margin m to the distributed examples on a hypersphere:

$$\mathcal{L}_{\text{clf}} = \tag{7}$$
$$\frac{-1}{N} \sum_{i=1}^{N} \log \frac{\exp(s \cos(\varphi_{y_i} + m))}{\exp(s \cos(\varphi_{y_i} + m)) + \sum_{\forall j \neq y_i} \exp(s \cos \varphi_j)},$$

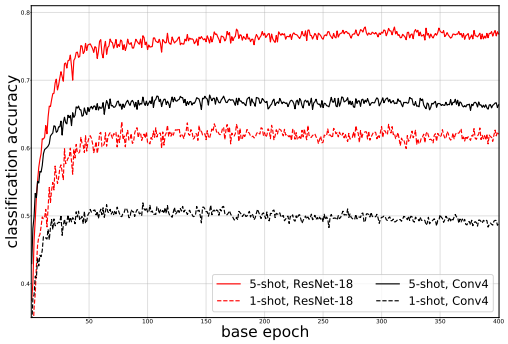


Figure 5: Validation error after fine-tuning as a function of the number of pre-training base epochs on *mini*-ImageNet with the cosmax loss function. Pre-training for a fixed number of iterations (here 400 as in [3]) may lead to overfitting the feature extraction on the base set. Each curve represents the average of 50 episodes.

where s is the radius of the hypersphere on which feature embeddings are distributed, N the number of examples, and m and s are hyperparameters (see sec. 6.1). The overall goal of the margin is to enforce inter-class discrepancy and intra-class compactness simultaneously.

6. Experimental validation

In the following, we are making an experimental evaluation and comparison of the proposed associative alignment strategies for few-shot learning. First, we introduce the datasets used and evaluate the strong baseline from sec. 7.

6.1. Datasets and implementation details

Datasets We used two benchmarks in our experiments: *mini*-ImageNet [33] (generic object recognition) and CUB-200-2011 (CUB) [34] (fine-grained image classification). *mini*-ImageNet is a subset of the ImageNet ILSVRC-12 dataset [27] containing 100 categories and 600 examples per class. We used the same splits as Ravi and Larochelle [25], where 64, 16, and 20 classes are used for the base, validation, and novel, respectively. The CUB dataset [34] contains 11,788 images from 200 bird categories. We used the same splits as Hilliard *et al.* [15] using 100, 50, and 50 classes for the base, validation, and novel classes, respectively.

Network architectures For the feature learner $f(\cdot|\theta)$, we experiment with both a 4-layer convolutional backbone ("Conv4") with input image resolution of 84×84, similar to [7, 25, 29]. We also experiment with a ResNet-18 [14] backbone with an input spatial resolution of 224×224. We use a single hidden layer MLP of 1024 dimensions as the critic network $h(\cdot|\phi)$ (c.f. sec. 4.1).

Table 1: Strong baseline evaluation, presenting the average 5-way classification accuracy of arcmax and early stopping compared to softmax ("baseline" in [3]) and cosmax ("baseline++" in [3]). Results are reported with a ResNet-18 backbone – see the appendix (A1) for Conv4 – with $\pm$ indicating the 95% confidence intervals over 600 episodes on *mini*-ImageNet.

Method	1-shot	5-shot
softmax [36]	51.75 ± 0.80	74.27 ± 0.63
softmax†	54.46 ± 0.85	75.62 ± 0.65
softmax$^{\dagger\diamond}$	56.63 ± 0.85	76.04 ± 0.60
cosmax [36]	51.87 ± 0.77	75.68 ± 0.63
cosmax†	54.21 ± 0.87	75.06 ± 0.61
cosmax$^{\dagger\diamond}$	54.58 ± 0.85	75.99 ± 0.57
arcmax†	57.54 ± 0.84	75.28 ± 0.61
arcmax$^{\dagger\diamond}$	$\mathbf{58.07} \pm 0.82$	$\mathbf{76.62} \pm 0.58$

† our implementation $^{\diamond}$ with early stopping

Implementation details Recall from sec. 3.1 that training consists of two stages: 1) pre-training using base categories $\mathcal{X}^b$; and 2) fine-tuning on novel categories $\mathcal{X}^n$. For pre-training, we use the early stopping algorithm from sec. 5.1 with a window size of 50. Standard data augmentation approaches (i.e., color jitter, random crops, left-right flips, as in [3]) have been employed, and the Adam algorithm with a learning rate of 10^{-3} and batch size of 64 is used for pre-training. The arcmax loss (eq. 7) is configured with $s = 20$ and $m = 0.1$ as hyperparameters values.

In the fine-tuning stage, episodes are defined by randomly selecting ($N = 5$) number of classes from the novel categories $\mathcal{X}^n$. k examples for each category are subsequently sampled ($k = 1$ and $k = 5$ in our experiments). As in Chen *et al.* [3], no standard data augmentation was used in this stage. We used episodic cross-validation to find s and m with a fixed encoder. More specifically, $s = 5$, and $m = 0.1$ or $m = 0.01$ are found for the Conv4 or ResNet-18 backbones, respectively. For the adversarial alignment procedure (sec. 4.1), the learning rate for the Adam algorithm was set to 10^{-5}. Similarly to [1], 5 iterations (inner loop of algorithm 1) were used to train the critic network $h(\cdot|\phi)$. We fix the number of related base categories as $B = 10$ (see appendix (A2) for Ablation study on B), as we observed that too few base examples could result in the feature learner forgetting what it already learned. For this reason, we used a relatively large number of categories (50 classes out of the 64 available in *mini*-ImageNet).

6.2. Baselines evaluation

Let first evaluate the new baseline presented in sec. 7. Table 5 show a comparison between using the softmax, cosmax,

Table 2: Few-shot classification results on *mini*-ImageNet using both Conv4 and ResNet-18 backbones. Our baseline is "arcmax$^{\dagger\diamond}$" from table 5 and $\pm$ denotes the 95% confidence intervals over 600 episodes.

		Conv4		ResNet-18	
		1-shot	5-shot	1-shot	5-shot
meta-learning	MatchingNet [33]	43.56 ± 0.84	55.31 ± 0.73	–	–
	MatchingNet‡ [33]	48.14 ± 0.78	63.48 ± 0.66	52.91 ± 0.88	68.88 ± 0.69
	ProtoNet [29]	49.42 ± 0.78	68.20 ± 0.66	–	–
	ProtoNet‡ [29]	44.42 ± 0.84	66.68 ± 0.69	54.16 ± 0.82	73.68 ± 0.65
	MAML [7]	48.07 ± 1.75	63.15 ± 0.91	–	–
	MAML‡ [7]	46.47 ± 0.82	62.71 ± 0.71	49.61 ± 0.92	65.72 ± 0.77
	SNAIL§ [22]	–	–	55.71 ± 0.99	68.88 ± 0.92
	TADAM§ [23]	–	–	58.5	76.70
	RelationNet [31]	50.44 ± 0.82	65.32 ± 0.70	–	–
	RelationNet‡ [31]	49.31 ± 0.85	66.60 ± 0.69	52.48 ± 0.86	69.83 ± 0.68
transfer learning	cosmax$^{\dagger\diamond}$ [3]	52.04 ± 0.82	68.47 ± 0.60	54.58 ± 0.85	75.99 ± 0.57
	our baseline (sec. 7)	51.90 ± 0.79	69.07 ± 0.62	58.07 ± 0.82	76.62 ± 0.58
alignment	adversarial	52.13 ± 0.99	70.78 ± 0.60	58.84 ± 0.77	77.92 ± 0.82
	centroid	**53.14** ± 1.06	**71.45** ± 0.72	**59.88** ± 0.67	**80.35** ± 0.73

$^{\dagger\diamond}$ our implementation, with early stopping ‡ implementation from [3] § ResNet-12 with additional learning module

and arcmax criteria. We report results from Chen *et al*. [3], but also those with our implementations of their method. The results show that arcmax with early stopping improves on using cosmax and softmax with and without early stopping for both the 1-shot and 5-shot, in 5-way scenarios. We followed the same dataset split configuration, network architecture, and implementation details given in [3] for our testing.

6.3. Visualizing the alignment strategies

Fig. 6 presents a 2D visualization of our adversarial and centroid alignment methods using t-SNE [20]. While both methods achieve similar results with $B = 1$, the centroid method results yields more discriminative class separation compared to the adversarial method with $B = 10$.

6.4. *mini*-ImageNet and CUB experiments

We now evaluate both of our proposed adversarial and centroid alignment methods (sec. 4) for few-shot learning, on both the *mini*-ImageNet and CUB datasets. We perform both 1- and 5-shot in 5-way for fine-tuning.

***mini*-ImageNet** Table 2 compares our proposed alignment methods with the strong baseline from sec. 7 and several state of the art approaches from the meta-learning and transfer learning literature including ProtoNet [29], MAML [7], RelationNet [31], TADAM [23], SNAIL [22], and cosmax [3].

While both alignment strategies offer improvements over the baseline, the centroid alignment outperforms all the methods. Indeed, it achieves the best 1-shot and 5-shot classification tasks on both the ResNet-18 and Conv4 backbones,

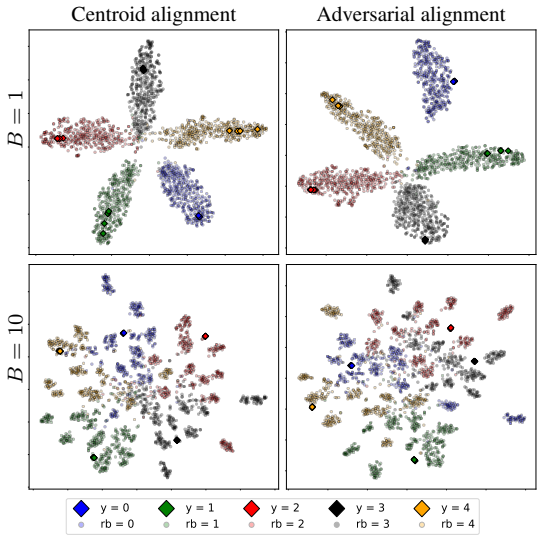


Figure 6: Aligning novel and related base classes. Columns present centroid and adversarial distribution matching while the rows compare picking $B = 1$ and $B = 10$ related base classes for each novel class. We use t-SNE [20] to visualize the 512-dimensional feature space of ResNet-18 in 2D. Results are for 5-shot in a 5-way setting.

with notable absolute accuracy improvements of 5.3% and 4.4% in 1- and 5-shot with the ResNet-18 backbone over the state of the art (i.e., cosmax [3]). It is important to note that methods such as SNAIL [22] and TADAM [23] make use of

Table 3: Few-shot classification results on the CUB dataset using both Conv4 and ResNet-18 backbones. Our baseline is "arcmax$^{\dagger\diamond}$" from table 5 and $\pm$ denotes the 95% confidence intervals over 600 episodes.

		Conv4		ResNet-18	
		1-shot	5-shot	1-shot	5-shot
meta-learning	ProtoNet‡ [29]	51.31 $\pm$ 0.91	70.77 $\pm$ 0.69	71.88 $\pm$ 0.91	87.42 $\pm$ 0.48
	MAML‡ [8]	55.92 $\pm$ 0.95	72.09 $\pm$ 0.76	69.96 $\pm$ 1.01	82.70 $\pm$ 0.65
	RelationNet‡ [31]	62.45 $\pm$ 0.98	76.11 $\pm$ 0.69	67.59 $\pm$ 1.02	82.75 $\pm$ 0.58
transfer learning	cosmax$^{\dagger\diamond}$ [3]	60.66 $\pm$ 1.04	79.79 $\pm$ 0.75	68.88 $\pm$ 0.92	84.37 $\pm$ 0.49
	our baseline (sec. 7)	60.85 $\pm$ 1.07	79.74 $\pm$ 0.64	71.71 $\pm$ 0.86	85.74 $\pm$ 0.49
alignment	adversarial	**63.30** $\pm$ 0.94	**81.35** $\pm$ 0.67	70.80 $\pm$ 1.12	88.04 $\pm$ 0.54
	centroid	62.71 $\pm$ 0.88	80.48 $\pm$ 0.81	**74.22** $\pm$ 1.09	**88.65** $\pm$ 0.55

$^{\dagger\diamond}$ our implementation, with early stopping ‡ implementation from [3]

extra modules such as dense and attention blocks [22] or task embedding networks [23]. In contrast, our method achieves significant improvements without changing the architecture.

CUB Table 3 similarly evaluates our approaches on the CUB dataset. As with *mini*-ImageNet, we observe improvements with both of our alignment approaches, with our centroid alignment outperforming the state of the art (ProtoNet [29]) by 2.3% and 1.2% with the ResNet-18 backbone in 1- and 5-shot, respectively. Here, the adversarial method slightly outperforms the centroid in the Conv4 case.

6.5. Cross-domain evaluation

We also evaluate our alignment strategies in cross-domain image classification. Here, following [3], the base categories are drawn from *mini*-ImageNet, during the novel categories from CUB. As shown in table 7, we gain 1.3% and 5.4% over the baseline in the 1- and 5-shot, respectively, with our proposed centroid alignment. Adversarial alignment falls below the baseline in 1-shot by -1.2%, but gains 5.9% in 5-shot. Overall, our centroid alignment method shows absolute accuracy improvements over the state of the art (i.e., cosmax [3]) of 3.8% and 6.0% in 1- and 5- shot respectively. One could argue that the three bird categories (i.e., house finch, robin, and toucan) in *mini*-ImageNet bias the cross-domain evaluation. Re-training the approach by excluding these classes resulted in a similar performance. Such quantitative results are given in the appendix (A3) materials.

7. Conclusion

In this paper, we have presented an associative alignment approach for the few-shot image classification. Our approach allows higher generalization performance by enabling the training of the entire network, still while avoiding overfitting. It does not require external or generated data, instead of leveraging base instances that are deemed similar to the novel examples, aligning them to support learning the same concepts. To do so, we apply a procedure to detect related base categories for each novel class. Then, we proposed two different strategies for keeping the intra-class alignment while performing updates for the classification task. Our experiments demonstrate that our approach, specifically the centroid-based alignment, significantly outperforms previous related works.

The current limitations of our work provide interesting future research directions are envisioned. First, the alignment approach (sec. 4) might include irrelevant intra-class examples from the base categories, so using categorical semantic information could help filter out bad base examples. Second, our training algorithms compute the related base examples once and subsequently keep them fixed during an episode, not taking into account the changes applied to the latent space during the episodic training. A more sophisticated dynamic sampling mechanism could be helpful in updating the examples which are the most helpful for training.

Table 4: Cross-domain image classification from *mini*-ImageNet to CUB in 1-shot and 5-shot scenarios using a ResNet-18 backbone. See appendix (A3) for 10-shot results.

	Method	1-shot	5-shot
meta-l.	ProtoNet‡ [36]	-	62.02 $\pm$ 0.70
	MAML‡ [8]	-	51.34 $\pm$ 0.72
	RelationNet‡ [31]	-	57.71 $\pm$ 0.73
tr. l.	cosmax$^{\dagger\diamond}$ [3]	43.06 $\pm$ 1.01	64.38 $\pm$ 0.86
	our baseline (sec. 7)	45.60 $\pm$ 0.94	64.93 $\pm$ 0.95
align.	adversarial	44.37 $\pm$ 0.94	**70.80** $\pm$ 0.83
	centroid	**46.85** $\pm$ 0.75	70.37 $\pm$ 1.02

$^{\dagger\diamond}$ our implementation, with early stopping
‡ implementation from [3]

Acknowledgement

This project was supported through funding from Mitacs, Prompt-Québec, NSERC-Canada, and E Machine Learning We thank Ihsen Hedhli, Marc-André Gardner, and Annette Schwerdtfeger for proofreading of the manuscript.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 4, 6

[2] 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

[3] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019. 1, 2, 3, 5, 6, 7, 8, 11

[4] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. Image deformation meta-networks for one-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[5] Wen-Hsuan Chu, Yu-Jhe Li, Jing-Cheng Chang, and Yu-Chiang Frank Wang. Spot and learn: A maximum-entropy patch sampler for few-shot image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018. 1, 5

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017. 1, 2, 5, 6, 7

[8] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, 2018. 2, 8

[9] Hang Gao, Zheng Shou, Alireza Zareian, Hanwang Zhang, and Shih-Fu Chang. Low-shot learning via covariance-preserving adversarial augmentation networks. In *Advances in Neural Information Processing Systems*, 2018. 2

[10] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*, 2017. 2

[11] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[12] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. *arXiv preprint arXiv:1905.01102*, 2019. 2

[13] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *IEEE International Conference on Computer Vision*, 2017. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, 2016. 6

[15] Nathan Hilliard, Lawrence Phillips, Scott Howland, Artëm Yankov, Courtney D Corley, and Nathan O Hodas. Few-shot learning with metric-agnostic conditional embeddings. *arXiv preprint arXiv:1802.04376*, 2018. 6

[16] Junsik Kim, Tae-Hyun Oh, Seokju Lee, Fei Pan, and In So Kweon. Variational prototyping-encoder: One-shot learning with prototypical images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[17] Taesup Kim, Jaesik Yoon, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. *arXiv preprint arXiv:1806.03836*, 2018. 2

[18] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 2

[19] Yann Lifchitz, Yannis Avrithis, Sylvaine Picard, and Andrei Bursuc. Dense classification and implanting for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[20] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008. 2, 7

[21] Akshay Mehrotra and Ambedkar Dukkipati. Generative adversarial residual pairwise networks for one shot learning. *arXiv preprint arXiv:1703.08033*, 2017. 2

[22] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017. 7, 8

[23] 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*, 2018. 2, 7, 8

[24] Hang Qi, Matthew Brown, and David G. Lowe. Low-shot learning with imprinted weights. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. 1, 2

[25] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016. 1, 2, 6

[26] 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. *arXiv preprint arXiv:1803.00676*, 2018. 2

[27] 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*, 2015. 6

[28] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition. In *Advances in Neural Information Processing Systems*, 2018. 2

[29] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 2017. 1, 2, 5, 6, 7, 8

[30] Qianru Sun, Xinzhe Li, Yaoyao Liu, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for

semi-supervised few-shot classification. *arXiv preprint arXiv:1906.00562*, 2019. 2

[31] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 7, 8

[32] Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 2002. 2

[33] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, 2016. 1, 2, 5, 6, 7

[34] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[35] Yong Wang, Xiao-Ming Wu, Qimai Li, Jiatao Gu, Wangmeng Xiang, Lei Zhang, and Victor OK Li. Large margin few-shot learning. *arXiv preprint arXiv:1807.02872*, 2018. 2

[36] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 6, 8, 11

[37] Davis Wertheimer and Bharath Hariharan. Few-shot learning with localization in realistic settings. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[38] Hongguang Zhang, Jing Zhang, and Piotr Koniusz. Few-shot learning via saliency-guided hallucination of samples. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 2

Appendix

A1 Strong baseline

Table 5 below complements Table 1 from the main paper by including results obtained with a Conv4 backbone (also commonly used in the literature). For completion, we also repeat the results obtained with ResNet-18 here. Arcmax plus early stopping outperforms the other baselines in all the other scenarios, except for the 1-shot scenario with Conv4.

Table 5: Strong baseline evaluation using both Conv4 and ResNet-18 backbone, presenting the average 5-way classification accuracy of arcmax and early stopping compared to softmax ("baseline" in [3]) and cosmax ("baseline++" in [3]). Results are reported with $\pm$ indicating the 95% confidence intervals over 600 episodes on *mini*-ImageNet.

	Conv4		ResNet-18	
	1-shot	5-shot	1-shot	5-shot
softmax [36]	42.11$\pm$ 0.71	62.53 $\pm$ 0.69	51.75 $\pm$ 0.80	74.27 $\pm$ 0.63
softmax†	46.40$\pm$ 0.72	64.37 $\pm$ 0.59	54.46 $\pm$ 0.85	75.62 $\pm$ 0.65
softmax$^{\dagger\diamond}$	46.99$\pm$ 0.73	65.33 $\pm$ 0.60	56.63 $\pm$ 0.85	76.04 $\pm$ 0.60
cosmax [36]	48.24 $\pm$ 0.75	66.43 $\pm$0.63	51.87 $\pm$ 0.77	75.68 $\pm$ 0.63
cosmax†	50.92 $\pm$ 0.76	67.29 $\pm$0.59	54.21 $\pm$ 0.87	75.06 $\pm$ 0.61
cosmax$^{\dagger\diamond}$	**52.04** $\pm$ 0.82	68.47 $\pm$0.60	54.58 $\pm$ 0.85	75.99 $\pm$ 0.57
arcmax†	51.79 $\pm$ 0.79	68.08 $\pm$ 0.64	57.54 $\pm$ 0.84	75.28 $\pm$ 0.61
arcmax$^{\dagger\diamond}$	51.90 $\pm$ 0.79	**69.07** $\pm$ 0.62	**58.07** $\pm$ 0.82	**76.62** $\pm$ 0.58

† our implementation $^{\diamond}$ with early stopping

A2 Ablation study on B

Table 6 presents an ablation study for B, the number of related base classes selected for each novel class. We perform the study on few-shot image classification on the *mini*-ImageNet dataset using ResNet-18 backbone. Overall, better results are obtained with a larger value of B, except for the adversarial alignment method in the 5-shot scenario.

Table 6: Effect of three different number of related bases B on few-shot classification results on *mini*-ImageNet using ResNet-18 backbones. $\pm$ denotes the 95% confidence intervals over 600 episodes.

B	1-shot	5-shot
1	55.76 $\pm$ 1.20	**79.34** $\pm$ 0.69
5	58.20 $\pm$ 1.14	78.65 $\pm$ 0.94
10	**58.84** $\pm$ 0.77	77.92 $\pm$ 0.82

(a) Adversarial alignment

B	1-shot	5-shot
1	58.04 $\pm$ 0.98	77.54 $\pm$ 0.73
5	58.97 $\pm$ 1.06	79.14 $\pm$ 0.91
10	**59.88** $\pm$ 0.67	**80.35** $\pm$ 0.73

(b) Centroid alignment

A3 Cross-domain

The *mini*-ImageNet dataset contains three birds classes (i.e., house finch, robin, and toucan) in the base categories, which might bias the cross-domain evaluation. Even though these three birds categories are not included in CUB dataset, we present experiments by excluding these classes from *mini*-ImageNet. We also include experiments in the 10-shot scenario in the table below.

Table 7: Additional cross-domain image classification from *mini*-ImageNet to CUB in 1-shot, 5-shot, and 10-shot scenarios using a ResNet-18 backbone to complement table 4 in the main paper. $\pm$ denotes the 95% confidence intervals over 600 episodes.

	Method	1-shot	5-shot	10-shot
	baseline	45.60 $\pm$ 0.94	64.93 $\pm$ 0.95	68.95$\pm$0.78
associative alignment	adversarial	44.37 $\pm$ 0.94	70.80 $\pm$ 0.83	79.63$\pm$0.71
	adversarial*	44.65 $\pm$ 0.88	71.48 $\pm$ 0.96	78.52$\pm$0.70
	centroid	46.85 $\pm$ 0.75	70.37 $\pm$ 1.02	79.98$\pm$0.80
	centroid*	47.25 $\pm$ 0.76	72.37 $\pm$ 0.89	79.46$\pm$0.72

* without birds (house finch, robin, toucan) in base classes

A4 Comparison to no alignment

Table 8 illustrates the effect of training the network using both novel and their related classes, but without the alignment losses. The results are shown in the "no alignment" row in table 8 below. Excluding the alignment loss slightly improves the accuracy compared to baseline by 0.82% and 0.24% in 1-shot and 5-shot using Conv4, respectively; however, it falls below the baseline by -2.13% and -2.34% in 1-shot and 5-shot using ResNet-18, respectively. In addition, except for the adversarial alignment in 1-shot using Conv4, both of the alignment strategies result accuracy improvement in all of the scenarios, which shows the necessity of an alignment strategy.

Table 8: Evaluating the necessity of alignment loss. Few-shot classification results on *mini*-ImageNet using both Conv4 and ResNet-18 backbones. $\pm$ denotes the 95% confidence intervals over 600 episodes.

		Conv4		ResNet-18	
		1-shot	5-shot	1-shot	5-shot
	baseline	51.90 $\pm$ 0.79	69.07 $\pm$ 0.62	58.07 $\pm$ 0.82	76.62 $\pm$ 0.58
	no alignment	52.72 $\pm$ 0.79	69.31 $\pm$ 0.69	55.94 $\pm$ 0.88	74.28 $\pm$ 0.83
alignment	adversarial	52.13 $\pm$ 0.99	70.78 $\pm$ 0.60	58.84 $\pm$ 0.77	77.92 $\pm$ 0.82
	centroid	**53.14** $\pm$ 1.06	**71.45** $\pm$ 0.72	**59.88** $\pm$ 0.67	**80.35** $\pm$ 0.73