Spatial Contrastive Learning for Few-Shot Classification

Yassine Ouali, Céline Hudelot, Myriam Tami

Spatial Contrastive Learning for Few-Shot Classification

Yassine Ouali Céline Hudelot Myriam Tami
Université Paris-Saclay, CentraleSupélec, MICS, 91190, Gif-sur-Yvette, France.
{yassine.ouali,celine.hudelot,myriam.tami}@centralesupelec.fr

Abstract

In this paper, we explore contrastive learning for few-shot classification, in which we propose to use it as an additional auxiliary training objective acting as a data-dependent regularizer to promote more general and transferable features. In particular, we present a novel attention-based spatial contrastive objective to learn locally discriminative and class-agnostic features. As a result, our approach overcomes some of the limitations of the cross-entropy loss, such as its excessive discrimination towards seen classes, which reduces the transferability of features to unseen classes. With extensive experiments, we show that the proposed method outperforms state-of-the-art approaches, confirming the importance of learning good and transferable embeddings for few-shot learning
Code: https://github.com/yassouali/SCL.

1. Introduction

Few-shot learning (Lake et al., 2011) has emerged as an alternative to supervised learning to simulate more realistic settings that mimic human capabilities, and in particular, it consists of reproducing the learner's ability to rapidly and efficiently adapt to novel tasks. In this paper, we tackle the problem of few-shot image classification, which aims to equip a learner with the ability to learn novel visual concepts and recognize unseen classes with limited supervision.

A popular paradigm to solve this problem is meta-learning (Thrun, 1998; Naik & Mammone, 1992) consisting of two disjoint stages, meta-training and meta-testing. During meta-training, the goal is to acquire transferable knowledge from a set of tasks sampled from the meta-training tasks so that the learner is equipped with the ability to adapt to novel tasks quickly. This fast adaptability to unseen classes is evaluated at test time by the average test accuracy over several meta-testing tasks. Such transferable knowledge can be acquired from the meta-training tasks with optimization-based methods (Ravi & Larochelle, 2017; Finn et al., 2017) or metric-based methods (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018).

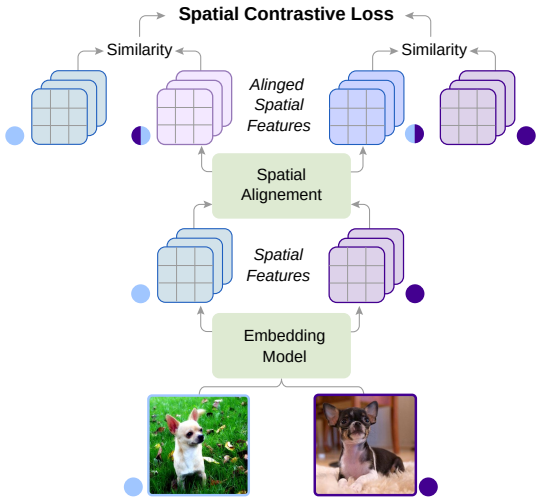

Figure 1. Spatial Contrastive Learning (SCL). To learn more locally class-independent discriminative features, we propose to measure the similarity between a given pair of samples using their spatial features as opposed to their global features. We first apply an attention-based alignment, aligning each input with respect to the other. Then, we measure the one-to-one spatial similarities and compute the Spatial Contrastive (SC) loss.

Recently, a growing line of works (Chen et al., 2019; Dhillon et al., 2020; Tian et al., 2020b) show that learning good representations results in fast adaptability at test time, suggesting that feature reuse (Raghu et al., 2019) plays a more important role in few-shot classification than the meta-learning aspect of existing algorithms. Such methods consider an extremely simple transfer learning baseline, in which the model is first pre-trained using the standard cross-entropy (CE) loss on the meta-training set. Then, at test time, a linear classifier is trained on the meta-testing set on top of the pre-trained model. The pre-trained model can either be fine-tuned (Dhillon et al., 2020; Afrasiyabi et al., 2020) together with the classifier, or fixed and used as a feature extractor (Chen et al., 2019; Tian et al., 2020b). While promising, we argue that using the CE loss during the pre-training stage hinders the quality of the learned representations since the model only acquires the necessary knowledge to solve the classification task over seen classes at train time. As a result, the learned visual features are excessively discriminative against the training classes, rendering them sub-optimal for

test time classification tasks constructed from an arbitrary set of unseen and novel classes.

To alleviate these limitations, we propose to leverage contrastive representation learning (Wu et al., 2018b; He et al., 2020; Chen et al., 2020) as an auxiliary objective, where instead of only mapping the inputs to fixed targets, we also optimize the features, pulling together semantically similar (*i.e.*, positive) samples in the embedding space while pushing apart dissimilar (*i.e.*, negative) samples. By integrating the contrastive loss into the learning objective, we give rise to discriminative representations between dissimilar instances while maintaining an invariance towards visual similarities. Subsequently, the learned representations are more transferable and capture more prevalent patterns outside of the seen classes. Additionally, by combining both losses, we leverage the stability of the CE loss and its effectiveness on small datasets and small batch sizes, while taking benefit of the contrastive loss as a data-dependent regularizer promoting more general-purpose embeddings. Additionally, by combining both losses, we leverage the stability of the CE loss and its effectiveness on small datasets and small batch sizes, in addition to taking benefit of the contrastive loss as a data-dependent regularizer promoting more general-purpose embeddings.

Specifically, we propose a novel attention-based spatial contrastive loss (see Fig. 1) as the auxiliary objective to further promote class-agnostic visual features and avoid suppressing local discriminative patterns. It consists of measuring the local similarity between the spatial features of a given pair of samples after an attention-based spatial alignment mechanism, instead of the global features (*i.e.*, avg. pooled spatial features) used in the standard contrastive loss. We also adopt the supervised formulation (Khosla et al., 2020) of the contrastive loss to leverage the provided label information when constructing the positive and negative samples.

However, directly optimizing the features and promoting the formation of clusters of similar instances in the embedding space might result in extremely disentangled representations. Such an outcome can be undesirable for few-shot learning, where the testing tasks can be notably different from the tasks encountered during training, *e.g.*, training on generic categories, and testing on fine-grained sub-categories. To solve this, we propose contrastive distillation to reduce the compactness of the features in the embedding space and provide additional refinement of the representations.

Contributions. To summarize, our contributions are:

- We explore contrastive learning as an auxiliary pre-training objective to learn more transferable features.
- We propose a novel Spatial Contrastive (SC) loss with an attention-based alignment mechanism to spatially compare a pair of features, further promoting class-independent discriminative patterns.
- We employ contrastive distillation to avoid excessive disentanglement of the learned embeddings and improve the performances.
- We demonstrate the effectiveness of the proposed method with extensive experiments on standard and cross-domain few-shot classification benchmarks, achieving state-of-the-art performances.
- We show the universality of the proposed method by applying it to a standard metric learning approach, resulting in a notable performance boost.

2. Preliminaries

Following a similar notation as (Tian et al., 2020b; Lee et al., 2019), we start by introducing the meta-learning formulation and the standard transfer learning baseline of (Tian et al., 2020b) in §2.1 and §2.2. Then, in §2.3, we analyze the quality of the learned features in such a setting, motivating the need for an alternative pre-training objective in order to learn more transferable embeddings.

2.1. Problem Definition

Few-shot classification usually involves a meta-training set $\mathcal{T}$ and a meta-testing set $\mathcal{S}$ with disjoint label spaces. The meta-training set discerns *seen* classes, while the meta-testing set discerns novel and *unseen* classes. Each one of the meta sets consists of a number of classification tasks where each task describes a pair of training (*i.e.*, support) and testing (*i.e.*, query) sets with few examples, *i.e.*, $\mathcal{T} = \{(\mathcal{D}_t^{\text{train}}, \mathcal{D}_t^{\text{test}})\}_{t=1}^{T}$ and $\mathcal{S} = \{(\mathcal{D}_q^{\text{train}}, \mathcal{D}_q^{\text{test}})\}_{q=1}^{Q}$, with each dataset containing pairs of images $\mathbf{x}$ and their ground-truth labels y.

The goal of few-shot classification is to learn a classifier f_θ parametrized by θ capable of exploiting the few training examples provided by the dataset $\mathcal{D}^{\text{train}}$ to correctly predict the labels of the test examples from $\mathcal{D}^{\text{test}}$ for a given task. However, given the high dimensionality of the inputs and the limited number of training examples, the classifier f_θ suffers from high variance. As such, the training and testing inputs are replaced with their corresponding features, which are produced by an embedding model f_ϕ parametrized by ϕ and then used as inputs to the classifier f_θ.

To this end, the objective of meta-training algorithms is to learn a good embedding model f_ϕ so that the average test error of the classifier f_θ is minimized. This usually involves two stages: first, a meta-training stage inferring the parameters ϕ of the embedding model using the meta-training set $\mathcal{T}$, followed by a meta-testing stage evaluating the embedding model's performance on meta-testing set $\mathcal{S}$.

Figure 2. Analysis of the Learned Representations. (a) k-Nearest Neighbors Analysis. For a given test image from *mini*-ImageNet dataset, we compute the nearest neighbors in the embedding space on the test set, and we observe that they are semantically dissimilar. This suggests that the learned embeddings are excessively discriminative towards features used to solve the training classification tasks (*e.g.*, the beer bottles in the second test image), which are not useful to recognize the novel classes at test time. (b) GradCAM results. To obtain the class activation maps (CAMs) explaining such an outcome, we train a linear classifier on the whole test set on top of the frozen embedding model and compute the CAMs. We see that the dominant discriminative features are not the ones useful for test-time classification.

2.2. Transfer Learning Baseline

In this work, we consider the simple transfer learning baseline of (Tian et al., 2020b), in which the embedding model f_ϕ is first pre-trained on the merged tasks from the meta-training set using the CE loss. Then, the model is carried over to the meta-testing stage and fixed during evaluation.

Concretely, we start by merging all the meta-training tasks $\mathcal{D}_t^{\text{train}}$ from $\mathcal{T}$ into a single training set $\mathcal{D}^{\text{new}}$ of seen classes:

$$\mathcal{D}^{\text{new}} = \cup\{\mathcal{D}_1^{\text{train}}, \dots, \mathcal{D}_t^{\text{train}}, \dots, \mathcal{D}_T^{\text{train}}\}. \quad (1)$$

Then, during the meta-training stage, the embedding model f_ϕ can be pre-trained on the resulting set of seen classes using the standard CE loss L_{CE}:

$$\phi = \arg\min_{\phi} L_{\text{CE}}(\mathcal{D}^{\text{new}}; \phi). \quad (2)$$

The pre-trained model f_ϕ is then fixed (*i.e.*, no fine-tuning is performed) and leveraged as a feature extractor during the meta-testing stage. For a given task $(\mathcal{D}_q^{\text{train}}, \mathcal{D}_q^{\text{test}})$ sampled from $\mathcal{S}$, a linear classifier f_θ is first trained on top of the extracted features to recognize the unseen classes using the training dataset $\mathcal{D}_q^{\text{train}}$:

$$\theta = \arg\min_{\theta} L_{\text{CE}}(\mathcal{D}_q^{\text{train}}; \theta, \phi) + \mathcal{R}(\theta), \quad (3)$$

where $\mathcal{R}$ is a regularization term, and the parameters $\theta = \{\mathbf{W}, \mathbf{b}\}$ consist of weight and bias terms, respectively. The predictor f_θ can then be used on the features of the test dataset $\mathcal{D}_q^{\text{test}}$ to obtain the class predictions and evaluate f_ϕ.

2.3. Analysis of the Learned Representations

Although the baseline of §2.2 delivers impressive results, we hypothesis that the usage of the CE loss during the meta-training stage can hinder the performances. Our intuition is that the learned representations lack general discriminative visual features since the CE loss induces embeddings tailored for solving the classification task over the seen classes. As a results, their transferability to novel domains with unseen classes is reduced, and especially if the domain gap between the training and testing stages is significant.

To empirically validate such a hypothesis, we conduct a k-nearest neighbor search (Johnson et al., 2017) on the learned embedding space. First, we train a model with the CE loss on the meta-training set of *mini*-ImageNet (Vinyals et al., 2016) as in Eq. (2). Then, for a given test image, we search for its neighbors from the meta-testing set. The results are shown in Fig. 2. For a fast test-time adaptation of the predictor f_θ, the desired outcome is to have visually and semantically similar images adjacent in the embedding space. However, we observe that the neighboring images are semantically dissimilar. Using Grad-CAM (Selvaraju et al., 2017), we notice that dominant discriminative features acquired during training might not be useful for discriminating between unseen classes at test time. In the case of *mini*-ImageNet, this observation is reinforced by the fact that the meta-training and meta-testing sets are closely related, in which better transferability of the learned features in expected when compared to other benchmarks. We note that similar behavior was also observed by (Doersch et al., 2020) for metric-learning based approaches.

To further investigate this behavior, we conduct a spectral analysis of the learned features. As shown in Fig. 3, we inspect the variance explained by a varying number of principal components and notice that almost all of the variance can be captured with a limited number of components, indicating that the CE loss only preserves the minimal amount of information required to solve the classification task. Similarity, by applying singular value decomposition to compute the eigen values of the feature matrix, we observe that the

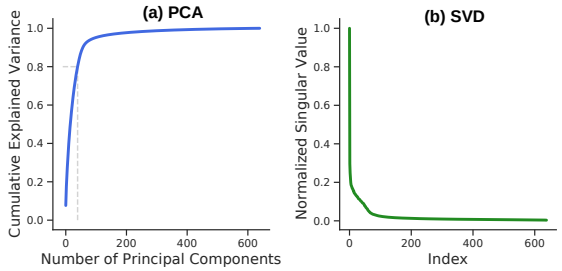

Figure 3. Spectral Analysis. Results of the spectral analysis on the embedding matrix. The plot (a) shows the explained cumulative variance of the learned features as the number of principal components used. We observe that 80% of the variance can be explained with only 30 components, indicating that embeddings lie in a lower dimensional space and are discriminative towards a small number of visual structures. Similarly, by computing the singular values of the embedding matrix, we see in (b) that the first singular values dominate the rest, indicating the same behavior.

maximal singular values are significantly larger than the remaining ones, diminishing the amount of informative signal that can be captured.

3. Methodology

Based on the observations presented in §2.3, in this section, we explore an alternative pre-training objective in order to learn a more transferable embedding model f_ϕ. First, in §3.1, we present the standard supervised contrastive loss. Then, in §3.2, we introduce a novel spatial contrastive learning objective followed by the pre-training objective in §3.3. Finally, an optional contrastive distillation step in §3.4.

3.1. Contrastive Learning

We explore contrastive learning as an auxiliary pre-training objective to learn general-purpose visual embeddings capturing discriminative features usable outside of the meta-training set. It thus facilitate the test time recognition of unseen classes. Specifically, given that in a few-shot classification setting we are provided with the class labels, we examine the usage of the supervised formulation (Khosla et al., 2020) of the contrastive loss which leverages the label information to construct the positive and negative samples.

Formally, let f_ϕ be an embedding model mapping the inputs $\mathbf{x}$ to *spatial* features $\mathbf{z}^{\mathrm{s}} \in \mathbb{R}^{HW \times d}$, followed by an average pooling operation to obtain the *global* features $\mathbf{z}^{\mathrm{g}} \in \mathbb{R}^{d}$, which are then mapped into a lower dimensional space using a projection head p, *i.e.*, $\mathbf{f} = p(\mathbf{z}^{\mathrm{g}})$ with $\mathbf{f} \in \mathbb{R}^{d'}$, and let a global similarity function $\mathrm{sim}_{\mathrm{g}}$ be denoted as the cosine similarity between a pair of projected global features $\mathbf{f}_i$ and $\mathbf{f}_j$ (*i.e.*, dot product between the ℓ_2 normalized features). First, we sample a batch of N pairs of images and labels from the merged meta-training set $\mathcal{D}^{\mathrm{new}}$ and augment each example in the batch, resulting in $2N$ data points. Then, the supervised contrastive loss (Khosla et al., 2020), referred to as the Global Contrastive (GC) loss, can be computed as follows:

$$L_{\mathrm{GC}} = \sum_{i=1}^{2N} \frac{1}{2N_{y_i} - 1} \sum_{j=1}^{2N} \mathbb{1}_{i \neq j} \cdot \mathbb{1}_{y_i = y_j} \cdot \ell_{ij} \,, \quad (4)$$

$$\text{where} \;\; \ell_{ij} = -\log \frac{\exp(\mathrm{sim}_{\mathrm{g}}(\mathbf{f}_i, \mathbf{f}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{i \neq k} \cdot \exp(\mathrm{sim}_{\mathrm{g}}(\mathbf{f}_i, \mathbf{f}_k)/\tau)} \,,$$

with $\mathbb{1}_{\mathrm{cond}} \in \{0, 1\}$ as an indicator function evaluating to 1 iff cond is satisfied, N_{y_i} as the total number of images with the same label y_i, and τ as a scalar temperature parameter. By using the GC loss of Eq. (4) as an additional pre-training objective with the CE loss, we push the embedding model f_ϕ to learn the visual similarities between instances of the same class, instead of only maintaining the useful features for the classification task over the seen classes, which results in more useful and transferable embeddings.

3.2. Spatial Contrastive Learning

Although the GC loss is capable of producing good embeddings, using the global features $\mathbf{z}^{\mathrm{g}}$ might suppress some local discriminative features present in the spatial features $\mathbf{z}^{\mathrm{s}}$ that can be informative at the meta-testing stage (*e.g.*, suppressing object specific features while overemphasizing the irrelevant background features). Additionally, encoding the relevant spatial information into the learned representations can play a critical role in increasing the robustness of the embeddings and reducing their sensitivity to domain changes, which is a highly desirable property for few-shot tasks. To this end, we propose a novel SC loss as an alternative objective, leveraging the spatial features $\mathbf{z}^{\mathrm{s}}$ to compute the similarity between a given pair of examples. However, to locally compare a pair of spatial features $\mathbf{z}^{\mathrm{s}}_i$ and $\mathbf{z}^{\mathrm{s}}_j$ and compute the SC loss, we first need to define a mechanism to align them spatially. To this end, we employ the attention mechanism (Vaswani et al., 2017) to compute the spatial attention weights to align the features $\mathbf{z}^{\mathrm{s}}_i$ with respect to $\mathbf{z}^{\mathrm{s}}_j$ and vice-versa. Then, we measure the one-to-one spatial similarity as illustrated in Fig. 4, and finally, compute the SC loss. a given pair of examples. However, to locally compare a pair of spatial features $\mathbf{z}^{\mathrm{s}}_i$ and $\mathbf{z}^{\mathrm{s}}_j$ and compute the SC loss, we first need to define a mechanism to align them spatially. To this end, we employ the attention mechanism (Vaswani et al., 2017) to compute the spatial attention weights to align the features $\mathbf{z}^{\mathrm{s}}_i$ with respect to $\mathbf{z}^{\mathrm{s}}_j$ and vice-versa. Then, we measure the one-to-one spatial similarity as illustrated in Fig. 4, and finally, compute the SC loss.

Attention-based Spatial Alignment. Let h_v, h_q and h_k denote the value, query and key projection heads, taking as input the spatial features $\mathbf{z}^{\mathrm{s}}$ and outputting the value $\mathbf{v}$, query $\mathbf{q}$ and key $\mathbf{k}$ of d'-dimensional features, *i.e.*, $\mathbf{v}, \mathbf{q}, \mathbf{k} \in \mathbb{R}^{HW \times d'}$. Given a pair of spatial features $\mathbf{z}^{\mathrm{s}}_i$ and $\mathbf{z}^{\mathrm{s}}_j$ of two

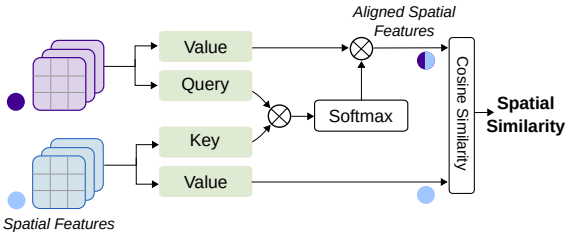

Figure 4. Attention-based Spatial Alignment. To compute the spatial similarity between a pair of features (purple and blue), we first spatially align the first features (purple) with respect to the second (blue) features with the attention mechanism (see Eq. (5)). Then we can compare the aligned value of the first features with the value of the second features. Note that the same process is applied in reverse to compute the final spatial similarity (see Eq. (6)).

instances i and j, we want to compute the aligned values of i with respect to j, denoted as $\mathbf{v}_{i|j}$. Such an alignment can be obtained using the key $\mathbf{k}_i$ and the query $\mathbf{q}_j$ to compute the attention weights $\mathbf{a}_{ij} \in \mathbb{R}^{HW \times HW}$, which can then be applied to $\mathbf{v}_i$ to obtain $\mathbf{v}_{i|j}$. Concretely, this can be computed as follows:

$$\mathbf{v}_{i|j} = \mathbf{a}_{ij}\mathbf{v}_i \;\text{ where }\; \mathbf{a}_{ij} = \text{softmax}\left(\frac{\mathbf{q}_j \mathbf{k}_i^\top}{\sqrt{d'}}\right). \quad (5)$$

Similarly, we compute $\mathbf{v}_{j|i}$ aligning the value of j with respect to i using the key $\mathbf{k}_j$ and the query $\mathbf{q}_i$.

Time Complexity. The spatial alignment mechanism has a time complexity of $O(N^2H^2W^2d'^2)$, which varies with the batch size, the size of the spatial features and the dimensionality of the values $\mathbf{v}$. To avoid excessive cost, for large input images, we apply an adaptive average pooling to reduce the size of the spatial features, in addition to using a small dimensionality d' and relatively small batches.

Spatial Similarity. Given a pair of values $\mathbf{v}_i$ and $\mathbf{v}_j$, together with their two aligned versions $\mathbf{v}_{i|j}$ and $\mathbf{v}_{j|i}$ computed using the attention mechanism detailed above, and with $\mathbf{v}_*^r$ denoting a feature vector at a spatial location $r \in [1, HW]$, we first perform an ℓ_2 normalization step of the values $\mathbf{v}_*^r$ at each spatial location r. Then, we compute the total spatial similarity $\text{sim}_\text{s}(\mathbf{z}_i^\text{s}, \mathbf{z}_j^\text{s})$ between a pair of spatial features as follows:

$$\text{sim}_\text{s}(\mathbf{z}_i^\text{s}, \mathbf{z}_j^\text{s}) = \frac{1}{HW}\sum_{r=1}^{HW}\left[(\mathbf{v}_i^r)^\top \mathbf{v}_{j|i}^r + (\mathbf{v}_j^r)^\top \mathbf{v}_{i|j}^r\right]. \quad (6)$$

Spatial Contrastive Learning. With the spatial similarity function sim_s defined in Eq. (6), and similar to the GC loss in Eq. (4), the SC loss can be computed as follows:

$$L_\text{SC} = \sum_{i=1}^{2N}\frac{1}{2N_{y_i}-1}\sum_{j=1}^{2N}\mathbb{1}_{i\neq j}\cdot\mathbb{1}_{y_i=y_j}\cdot\ell_{ij}\,, \quad (7)$$

$$\text{where}\;\; \ell_{ij} = -\log\frac{\exp(\text{sim}_\text{s}(\mathbf{z}_i^\text{s}, \mathbf{z}_j^\text{s})/\tau')}{\sum_{k=1}^{2N}\mathbb{1}_{i\neq k}\cdot\exp(\text{sim}_\text{s}(\mathbf{z}_i^\text{s}, \mathbf{z}_k^\text{s})/\tau')}\,,$$

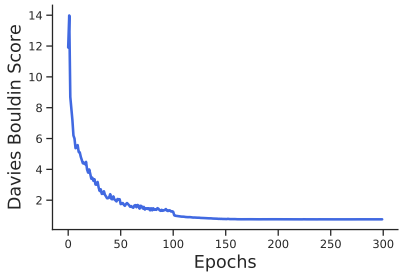

Figure 5. Degree of Clustering. The plot shows the evolution of the intra-class variation using the Davies-Bouldin index (Davies & Bouldin, 1979) during the course of training on *mini*-ImageNet when using the contrastive loss. We see that the learned embeddings of each class are significantly over-clustered, an outcome that might not be desired in some cases.

with τ' as a scalar temperature parameter.

3.3. Pre-training Objective

Based on the contrastive objectives in Eq. (4) and Eq. (7), the pre-training objective can take different forms. We mainly consider the case where the pre-training objective L_T is the summation of the CE and SC losses, with λ_CE and λ_SC as scaling weights to control the contribution of each term:

$$L_\text{T} = \lambda_\text{CE}L_\text{CE} + \lambda_\text{SC}L_\text{SC}. \quad (8)$$

However, we also explore other alternatives such as replacing L_SC with L_GC or training with both L_GC and L_SC as auxiliary losses with their corresponding weighting terms. Additionally, we also consider the self-supervised formulations of the GC and SC losses, where the label information is discarded and the only positives considered are the augmented versions of each example (*i.e.*, $y_i = i \mod N$). We refer to them as SS-GC and SS-SC (Self-Supervised Global and Spatial Contrastive) losses respectively.

Using the total loss L_T, the embedding model f_ϕ can be trained together with the projection head and the attention modules during the meta-training stage. Specifically, let ψ represent the parameters of the projection head p and the attention modules h_v, h_q and h_k. The parameters are obtained as follows:

$$\{\phi, \psi\} = \underset{\{\phi,\psi\}}{\arg\min}\, L_\text{T}(\mathcal{D}^\text{new}; \{\phi, \psi\}). \quad (9)$$

After the pre-training stage, the parameters ψ are discarded, and the embedding model f_ϕ is then fixed and carried over from meta-training to meta-testing.

3.4. Avoiding Excessive Disentanglement

Since the contrastive objectives encourage closely aligned embeddings of instances of the same class while distributing all of the normalized features uniformly on the hypersphere

(Wang & Isola, 2020), we have to consider a possible over-clustering of the features of the same class (see Fig. 5). Such an outcome can be desired for closed-set recognition, but in a few-shot setting, in which the discrepancy between the meta-training and meta-testing domain might differ greatly from one case to the other (*e.g.*, training on coarse seen categories, and testing on fine-grained unseen sub-categories), this might lead to sub-optimal performances. As such, to avoid an excessive disentanglement of the learned features and to further improve the generalization of the embedding model, we propose Contrastive Distillation (CD) to reduce the compactness of the features in embeddings space.

Contrastive Distillation. Given a teacher model f_{ϕ_t} pre-trained with the objective in Eq. (8), we transfer its knowledge to a student model f_{ϕ_s} using the standard knowledge distillation (Hinton et al., 2015) objective L_{KL} (*i.e.*, the Kullback-Leibler (KL) divergence between the student's predictions and the soft targets predicted by the teacher), but with an additional contrastive distillation loss L_{CD}. This loss consists of maximizing the inner dot product between the ℓ_2 normalized global features of the teacher $\mathbf{z}^{gt}$ and that of the student $\mathbf{z}^{gs}$, which corresponds to minimizing the squared Euclidean distance, formally:

$$L_{\text{CD}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{z}_i^{\text{g}t} - \mathbf{z}_i^{\text{g}s}\|_2^2. \tag{10}$$

To summarize, the student's parameters are learned as follows:

$$\phi_s = \arg\min_{\phi_s} \lambda_{\text{CD}} L_{\text{CD}}(\mathcal{D}^{\text{new}}; \phi_s, \phi_t) + \lambda_{\text{KL}} L_{\text{KL}}(\mathcal{D}^{\text{new}}; \phi_s, \phi_t). \tag{11}$$

As result, and different from the standard contrastive distillation loss (Tian et al., 2019b) that leverages negative samples, only maximizing the similarity between the pairs of features without using any negatives relaxes the uniformity constraint of the contrastive loss, which in turns reduces the disentanglement of the learned embeddings.

4. Experiments

For the experimental section, we base our implementation on the publicly available code of (Tian et al., 2020b) and conduct experiments on four popular few-shot classification benchmarks: *mini*-ImageNet (Vinyals et al., 2016), *tiered*-ImageNet (Ren et al., 2018), CIFAR-CS (Bertinetto et al., 2019) and FC100 (Oreshkin et al., 2018). Additionally, we present experiments on cross-domain few-shot benchmarks introduced by (Tseng et al., 2020). We note that additional experimental details and results are presented in the supplementary material.

4.1. Experimental Details

Architecture. For the embedding model f_ϕ, we follow (Tian et al., 2020b) and use a ResNet-12 consisting of 4 residual blocks with Dropblock as a regularizer and 640-dimensional output features (*i.e.*, $d = 640$). For the projection head and the attention modules, we use an MLP with one hidden layer and a ReLU non-linearity similar to SimCLR, outputting 80-dimensional features (*i.e.*, $d' = 80$).

Training Setup. For optimization, we use SGD with a momentum of 0.9, a weight decay of 5×10^{-4}, a learning rate of 5×10^{-2} and a batch size of 64. For the loss functions, we set the temperature parameters τ and τ' to 0.1 and the scaling weights λ_{CE}, λ_{SC}, and λ_{GC} to 1.0, except for CIFAR-FS where we set them to 0.5. For distillation, we set λ_{CD} to 10.0 and λ_{KL} to 1.0 and use a temperature of 4.0 for the KL loss.

Data Augmentation. During meta-training, for a given augmented batch of $2N$ examples, and consistent with other approaches (Tian et al., 2020b; Lee et al., 2019), the first N instances are obtained using standard augmentations, *i.e.*, random crop, color jittering and random horizontal flip. The remaining N instances are obtained with SimCLR type augmentations, *i.e.*, random resized crop, color jittering, random horizontal flip and random grayscale conversion. During the meta-testing stage, we follow (Tian et al., 2020b) and create 5 augmented versions of each training image to overcome the problem of data insufficiency and train the linear classifier f_θ.

Evaluation Setup. During meta-testing, and given a pre-trained embedding model f_ϕ, we follow (Tian et al., 2020b) and consider a linear classifier as the predictor f_θ, implemented in scikit-learn (Pedregosa et al., 2011) and trained on the ℓ_2 normalized features produced by f_ϕ. Specifically, we sample a number of C-way K-shot testing classification tasks constructed from the unseen classes of the meta-testing set, with C as the number of classes and K as the number of training examples per class. After training f_θ on the train set, the predictor is then applied to the features of the test set to obtain the prediction and compute the accuracy. In our case, we evaluate the model over 600 randomly sampled tasks and report the median accuracy over 3 runs with 95% confidence intervals, where in each run, the accuracy is the mean accuracy of the 600 sampled tasks.

4.2. Ablation Studies

We start by conducting detailed ablation studies to analyze the contribution of each component of the proposed method, from the choices of the loss function to the hyperparameters of the SC loss.

Loss Functions. To investigate the effect of the contrastive losses when used as auxiliary training objectives, we evalu-

Loss Function	Aug.	*mini*-ImageNet, 5-way 1-shot	*mini*-ImageNet, 5-way 5-shot	CIFAR-CS, 5-way 1-shot	CIFAR-CS, 5-way 5-shot
CE		61.8 ± 0.7	79.7 ± 0.6	71.3 ± 0.9	86.1 ± 0.6
CE	✓	61.8 ± 0.8	78.6 ± 0.5	71.9 ± 0.9	86.3 ± 0.5
CE + SS-GC	✓	62.7 ± 0.7	81.0 ± 0.6	70.9 ± 0.9	84.5 ± 0.6
CE + SS-SC	✓	64.0 ± 0.8	81.5 ± 0.5	72.1 ± 0.8	86.2 ± 0.6
CE + SS-GC + SS-SC	✓	62.8 ± 0.8	81.1 ± 0.6	69.0 ± 0.9	85.0 ± 0.6
CE + GC	✓	65.0 ± 0.8	81.6 ± 0.5	74.0 ± 0.8	87.3 ± 0.6
CE + SC	✓	$\mathbf{65.7 \pm 0.8}$	$\mathbf{82.5 \pm 0.5}$	75.0 ± 0.9	87.4 ± 0.6
CE + GC + SC	✓	65.0 ± 0.8	81.3 ± 0.5	$\mathbf{76.0 \pm 0.7}$	$\mathbf{87.5 \pm 0.5}$

Table 1. Comparison of the mean acc. obtained on *mini*-ImageNet and CIFAR-FS with different training objectives. "Aug." indicates the usage of SimCLR type augmentations.

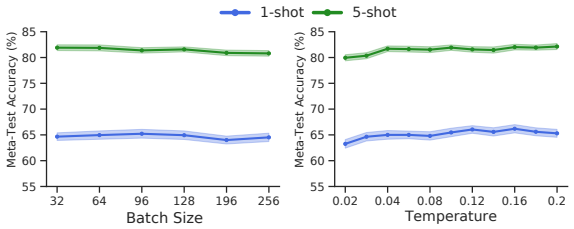

Figure 6. Comparison of the mean acc. obtained on *mini*-ImageNet across various batch sizes and SC loss temperatures.

Augmentation	*mini*-ImageNet, 5-way 1-shot	*mini*-ImageNet, 5-way 5-shot	CIFAR-CS, 5-way 1-shot	CIFAR-CS, 5-way 5-shot
Standard	64.3 ± 0.7	80.6 ± 0.5	74.9 ± 0.8	86.3 ± 0.6
SimCLR	$\mathbf{65.7 \pm 0.8}$	$\mathbf{82.5 \pm 0.5}$	$\mathbf{75.0 \pm 0.9}$	87.4 ± 0.6
AutoAugment	65.2 ± 0.7	82.1 ± 0.5	74.0 ± 0.9	86.7 ± 0.6
Stacked RandAug.	64.9 ± 0.8	81.6 ± 0.6	$\mathbf{75.0 \pm 0.9}$	$\mathbf{87.6 \pm 0.6}$

Table 2. Comparison of the mean acc. obtained on *mini*-ImageNet and CIFAR-FS with different augmentation strategies, which are used to obtain the additional N augmented instances within a minibatch.

Aggregation	*mini*-ImageNet, 5-way 1-shot	*mini*-ImageNet, 5-way 5-shot	CIFAR-CS, 5-way 1-shot	CIFAR-CS, 5-way 5-shot
Sum	65.2 ± 0.8	81.2 ± 0.5	$\mathbf{75.3 \pm 0.8}$	87.3 ± 0.5
Mean	$\mathbf{65.7 \pm 0.8}$	$\mathbf{82.5 \pm 0.5}$	75.0 ± 0.9	$\mathbf{87.4 \pm 0.6}$
Maximum	65.5 ± 0.7	82.0 ± 0.5	73.4 ± 0.8	86.4 ± 0.6
LogSumExp	64.8 ± 0.8	81.7 ± 0.6	74.2 ± 0.8	87.0 ± 0.6

Table 3. Comparison of the mean acc. obtained on *mini*-ImageNet and CIFAR-FS with different aggregation functions, which are used to amount the total similarity from the one-to-one spatial similarities.

Features Used	*mini*-ImageNet, 5-way 1-shot	*mini*-ImageNet, 5-way 5-shot	CIFAR-CS, 5-way 1-shot	CIFAR-CS, 5-way 5-shot
Spatial	64.5 ± 0.8	82.1 ± 0.5	$\mathbf{75.0 \pm 0.9}$	87.1 ± 0.6
Global	65.7 ± 0.8	82.5 ± 0.5	$\mathbf{75.0 \pm 0.9}$	87.4 ± 0.6
Glo. & Spa. (Max)	65.6 ± 0.8	82.1 ± 0.5	74.2 ± 0.8	87.3 ± 0.5
Glo. & Spa. (Sum)	$\mathbf{65.7 \pm 0.8}$	$\mathbf{83.1 \pm 0.5}$	$\mathbf{75.6 \pm 0.9}$	$\mathbf{87.6 \pm 0.6}$

Table 4. Comparison of the mean acc. obtained on *mini*-ImageNet and CIFAR-FS with different evaluation settings, in which we use either the global features, the spatial features, or both.

ate the performances obtained with various loss functions as detailed in §3.3. The results are shown in Table 1. We observe a notable gain in performance when adopting auxiliary contrastive losses, be it supervised or self-supervised,

Loss Function	*mini*-ImageNet, 5-way 1-shot	*mini*-ImageNet, 5-way 5-shot	CIFAR-CS, 5-way 1-shot	CIFAR-CS, 5-way 5-shot
Teacher	*65.7 ± 0.8*	*82.5 ± 0.5*	*75.0 ± 0.9*	*87.4 ± 0.6*
KL	66.0 ± 0.8	82.5 ± 0.5	75.9 ± 0.9	87.4 ± 0.6
KL+CD	$\mathbf{67.4 \pm 0.8}$	$\mathbf{82.7 \pm 0.5}$	$\mathbf{76.5 \pm 0.9}$	$\mathbf{87.6 \pm 0.6}$

Table 5. Comparison of the mean acc. obtained on *mini*-ImageNet and CIFAR-FS with different distillation objectives.

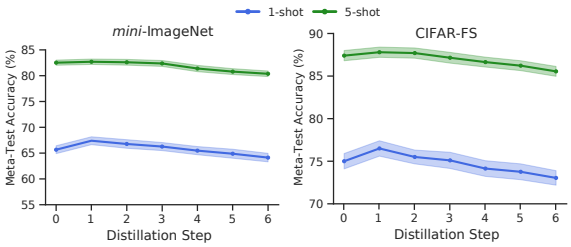

Figure 7. Comparison of the mean acc. obtained on *mini*-ImageNet and CIFAR-FS with sequential distillation.

with better gains when using the supervised formulation, highlighting the benefits of using the label information when constructing the positives and negatives samples. More importantly, the SC loss outperforms the standard GC loss, confirming the effectiveness of using the spatial features rather than the global features. Additionally, using both the SC and GC losses does not result in distinct gains over the SC loss. Thus, for the rest of this section, we adopt the SC as a sole auxiliary loss.

Spatial Contrastive Loss. In this section, we examine different variations and hyperparameters of the SC loss when used as an auxiliary objective along the CE loss. In particular, we consider the following variations:

- *Hyperparameters.* To inspect the SC loss's hyperparameter stability, we conduct experiments with different batch sizes and temperature values. As seen in Fig. 6, by combining the CE and the SC losses, we leverage the stability of the CE and obtain consistent results across several batch sizes, circumventing the need for very large batches when training with only the contrastive losses as it is the case in the unsupervised representation learning setting. As for the temperatures, disregarding the low temperatures in which the SC loss is dominated by the small distances, rendering the actual distances between widely separated representations almost irrelevant, we see comparable performances for temperatures above 0.05, further confirming the stability of the approach.

- *Augmentations.* Although we mainly use SimCLR type augmentations to produce the additional N augmented examples within a given batch, other augmentations can also be used. Specifically, we consider the standard augmentations used when training with only the CE loss, AutoAugment (Cubuk et al., 2019) and Stacked RandAugment (Tian et al., 2020a). Table 2 shows that the SimCLR type augmentations yield the best results overall. We speculate that

Method	Backbone	*mini*-ImageNet, 5-way 1-shot	*mini*-ImageNet, 5-way 5-shot	*tiered*-ImageNet, 5-way 1-shot	*tiered*-ImageNet, 5-way 5-shot
MAML (Finn et al., 2017)	32-32-32-32	48.70 ± 1.84	63.11 ± 0.92	51.67 ± 1.81	70.30 ± 1.75
Matching Networks (Vinyals et al., 2016)	64-64-64-64	43.56 ± 0.84	55.31 ± 0.73	-	-
Prototypical Networks† (Snell et al., 2017)	64-64-64-64	49.42 ± 0.78	68.20 ± 0.66	53.31 ± 0.89	72.69 ± 0.74
Relation Networks (Sung et al., 2018)	64-96-128-256	50.44 ± 0.82	65.32 ± 0.70	54.48 ± 0.93	71.32 ± 0.78
SNAIL (Mishra et al., 2018)	ResNet-12	55.71 ± 0.99	68.88 ± 0.92	-	-
TADAM (Oreshkin et al., 2018)	ResNet-12	58.50 ± 0.30	76.70 ± 0.30	-	-
Shot-Free (Ravichandran et al., 2019)	ResNet-12	59.04 ± n/a	77.64 ± n/a	63.52 ± n/a	82.59 ± n/a
MetaOptNet (Lee et al., 2019)	ResNet-12	62.64 ± 0.61	78.63 ± 0.46	65.99 ± 0.72	81.56 ± 0.53
Diversity w/ Coop. (Dvornik et al., 2019)	ResNet-18	59.48 ± 0.65	75.62 ± 0.48	-	-
Boosting (Gidaris et al., 2019)	WRN-28-10	63.77 ± 0.45	80.70 ± 0.33	70.53 ± 0.51	84.98 ± 0.36
Fine-tuning (Dhillon et al., 2020)	WRN-28-10	57.73 ± 0.62	78.17 ± 0.49	66.58 ± 0.70	85.55 ± 0.48
LEO-trainval† (Rusu et al., 2019)	WRN-28-10	61.76 ± 0.08	77.59 ± 0.12	66.33 ± 0.05	81.44 ± 0.09
RFS (Tian et al., 2020b)	ResNet-12	62.02 ± 0.63	79.64 ± 0.44	69.74 ± 0.72	84.41 ± 0.55
RFS-Distill (Tian et al., 2020b)	ResNet-12	64.82 ± 0.60	82.14 ± 0.43	71.52 ± 0.69	86.03 ± 0.49
Ours	ResNet-12	65.69 ± 0.81	83.10 ± 0.52	71.48 ± 0.89	**86.88 ± 0.53**
Ours-Distill	ResNet-12	**67.40 ± 0.76**	**83.19 ± 0.54**	**71.98 ± 0.91**	86.19 ± 0.59

Table 6. Comparison with prior few-shot classification works on ImageNet derivatives. We show the mean acc. and 95% confidence interval. †results obtained by training on both train and validation sets.

Method	Backbone	CIFAR-FS, 5-way 1-shot	CIFAR-FS, 5-way 5-shot	FC100, 5-way 1-shot	FC100, 5-way 5-shot
MAML (Finn et al., 2017)	32-32-32-32	58.9 ± 1.9	71.5 ± 1.0	-	-
Relation Networks (Sung et al., 2018)	64-96-128-256	55.0 ± 1.0	69.3 ± 0.8	-	-
R2D2 (Bertinetto et al., 2019)	96-192-384-512	65.3 ± 0.2	79.4 ± 0.1	-	-
TADAM (Oreshkin et al., 2018)	ResNet-12	-	-	40.1 ± 0.4	56.1 ± 0.4
Shot-Free (Ravichandran et al., 2019)	ResNet-12	69.2 ± n/a	84.7 ± n/a	-	-
TEWAM (Qiao et al., 2019)	ResNet-12	70.4 ± n/a	81.3 ± n/a	-	-
Prototypical Networks† (Snell et al., 2017)	ResNet-12	72.2 ± 0.7	83.5 ± 0.5	37.5 ± 0.6	52.5 ± 0.6
Boosting (Gidaris et al., 2019)	WRN-28-10	73.6 ± 0.3	86.0 ± 0.2	-	-
MetaOptNet (Lee et al., 2019)	ResNet-12	72.6 ± 0.7	84.3 ± 0.5	41.1 ± 0.6	55.5 ± 0.6
RFS (Tian et al., 2020b)	ResNet-12	71.5 ± 0.8	86.0 ± 0.5	42.6 ± 0.7	59.1 ± 0.6
RFS-Distill (Tian et al., 2020b)	ResNet-12	73.9 ± 0.8	86.9 ± 0.5	44.6 ± 0.7	60.9 ± 0.6
Ours	ResNet-12	75.6 ± 0.9	87.6 ± 0.6	44.4 ± 0.8	60.8 ± 0.8
Ours-Distill	ResNet-12	**76.5 ± 0.9**	**88.0 ± 0.6**	**44.8 ± 0.7**	**61.4 ± 0.7**

Table 7. Comparison with prior few-shot classification works on CIFAR-10 derivatives. We show the mean acc. and 95% confidence interval. †results obtained by training on both train and validation sets.

Method	*mini*-ImageNet 20-way 1-shot
Matching networks (Vinyals et al., 2016)	17.31 ± 0.22
Meta-LSTM (Ravi & Larochelle, 2017)	16.70 ± 0.23
MAML (Finn et al., 2017)	16.49 ± 0.58
Meta-SGD (Li et al., 2017)	17.56 ± 0.64
LGM-Net (Li et al., 2019)	26.14 ± 0.34
Ours	36.81 ± 0.38
Ours-Distill	**37.47 ± 0.32**

Table 8. Comparison with prior few-shot classification works on 20-way 1-shot *mini*ImageNet classification. We show the mean acc. and 95% confidence interval.

for the standard augmentation, without any novel transformations that the model is forced to be invariant under, the gains are minimal. As for strong augmentations (*i.e.*, AutoAugment and Stacked RandAugment), the augmented inputs might be substantially deformed, making the spatial alignment insufficient and reducing the effect of the SC loss.

- *Aggregation Function.* Table 3 presents the results obtained with various aggregation functions used to aggregate the one-to-one spatial similarities into an overall measure. We observe that when using the mean as the aggregate, we obtain overall better performances across the different datasets and settings.

Distillation.

To improve the generalization of the embedding model, we investigate the effect of knowledge distillation by training a new (*i.e.*, student) model using a pre-trained (*i.e.*, teacher) network with various training objectives. Table 5 shows a clear performance gain with the proposed CD objective as an additional loss term, confirming the benefits of optimizing the learned features and relaxing the compactness of the embedding space.

Additionally, we explore sequential self-distillation similar to Born-again networks (Furlanello et al., 2018), where we consider the student model as the teacher and repeat the distillation process. As detailed in Fig. 7, we notice a clear drop in performances beyond a single distillation step. We suspect this might be a result of an over disentanglement of the features induced by the CD loss. As such, for the rest of

Method	CUB, 5-way		Cars, 5-way		Places, 5-way		Plantae, 5-way	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MatchingNet (Vinyals et al., 2016)	35.89 ± 0.5	51.37 ± 0.7	30.77 ± 0.5	38.99 ± 0.6	49.86 ± 0.8	63.16 ± 0.8	32.70 ± 0.6	46.53 ± 0.6
MatchingNet w/ FT (Tseng et al., 2020)	36.61 ± 0.6	55.23 ± 0.8	29.82 ± 0.4	41.24 ± 0.6	51.07 ± 0.7	64.55 ± 0.7	34.48 ± 0.5	41.69 ± 0.6
RelationNet (Sung et al., 2018)	42.44 ± 0.7	57.77 ± 0.7	29.11 ± 0.6	37.33 ± 0.7	48.64 ± 0.8	63.32 ± 0.8	33.17 ± 0.6	44.00 ± 0.6
RelationNet w/ FT (Tseng et al., 2020)	44.07 ± 0.7	59.46 ± 0.7	28.63 ± 0.6	39.91 ± 0.7	50.68 ± 0.9	66.28 ± 0.7	33.14 ± 0.6	45.08 ± 0.6
GNN (Garcia & Bruna, 2018)	45.69 ± 0.7	62.25 ± 0.6	31.79 ± 0.5	44.28 ± 0.6	53.10 ± 0.8	70.84 ± 0.6	35.60 ± 0.5	52.53 ± 0.6
GNN w/ FT (Tseng et al., 2020)	47.47 ± 0.6	66.98 ± 0.7	31.61 ± 0.5	44.90 ± 0.6	55.77 ± 0.8	73.94 ± 0.7	35.95 ± 0.5	53.85 ± 0.6
Ours	49.58 ± 0.7	67.64 ± 0.7	34.46 ± 0.6	$\mathbf{52.22 \pm 0.7}$	59.37 ± 0.7	76.46 ± 0.6	$\mathbf{40.23 \pm 0.6}$	59.38 ± 0.6
Ours-Distill	$\mathbf{50.09 \pm 0.7}$	$\mathbf{68.81 \pm 0.6}$	$\mathbf{34.93 \pm 0.6}$	51.72 ± 0.7	$\mathbf{60.32 \pm 0.8}$	$\mathbf{76.51 \pm 0.6}$	39.75 ± 0.8	$\mathbf{59.91 \pm 0.6}$

Table 9. Comparison with prior works on cross-domain few-shot classification benchmarks. We train the model on the *mini*-ImageNet domain and evaluate the trained model on other domains. We show the mean acc. and 95% confidence interval.

the paper, we only apply a single distillation step to refine the features further while preserving the learned structures.

Evaluation. Up until now, we primarily trained a linear classifier on top of the global features during the meta-testing stage. Nonetheless, given that we explicitly optimize the spatial features during training, which increases their discriminability, we investigate their usage as inputs to the linear classifier. To this end, we compare the performance when training over the global features, the spatial features, or both, where we train two classifiers and aggregate their predictions. Table 4 shows the evaluation results. Overall, using the global features to train the linear classifier offers slightly better results than the spatial features. We suspect this might result from slight overfitting of the classifier given that the spatial features increase the number of parameters to be learned, which negatively impacts the performances. However, when leveraging both the spatial and global features, we obtain better results confirming the usefulness of the spatial feature even during the meta-testing stage.

4.3. Few-shot Classification

In this section, and based on the results of the ablation studies, we fix the training objective as SC+CE during the meta-training stage and use both the spatial and global feature during the meta-testing stage with a sum aggregate, and compare our approach with other popular few-shot classification methods.

ImageNet derivatives. The *mini*-ImageNet benchmark is a standard dataset used for few-shot image classification. It consists of 100 randomly selected classes from ImageNet (Russakovsky et al., 2015). Following (Ravi & Larochelle, 2017), the classes are split into 64, 16, and 20 classes for meta-training, meta-validation, and meta-testing respectively. Each class contains 600 images of size 84×84. The *tiered*-ImageNet (Ren et al., 2018) benchmark presents a larger subset of ImageNet, with 608 classes and images of size 84×84 assembled into 34 super-categories. These are split into 20 categories for meta-training and 6 categories for both meta-validation and meta-testing aiming to minimize the semantic similarity between the split.

CIFAR derivatives. CIFAR-CS (Bertinetto et al., 2019) and FC100 (Oreshkin et al., 2018) are both CIFAR-100 (Krizhevsky et al., 2010) derivatives, containing 100 classes and images of size 32×32. For CIFAR-CS, the classes are divided into 64, 16 and 20 classes for meta-training, meta-validation, and meta-testing respectively. As for FC100, the classes are grouped into 20 super-categories, which are split into 12 categories for meta-training and 4 categories for both meta-validation and meta-testing.

Results. The results of 5-way classification are summarized in Table 6 and Table 7 for ImageNet and CIFAR derivatives respectively, in addition to 20-way classification results on *mini*-ImageNet in Table 8. Our method outperforms previous works and achieves state-of-the-art performances across different datasets and evaluation settings. This suggests that our attention-based SCL approach coupled with the CE loss improves the transferability of the learned embeddings without any meta-learning techniques, with additional improvements using a contrastive distillation step. These results also show the potential of integrating contrastive losses as auxiliary objectives for various few-shot learning scenarios.

4.4. Cross-Domain Few-shot Classification

To further affirm the improved transferability of the learned embedding with our approach, we explore the effects of an increased domain difference between the seen and unseen classes, *i.e.*, the discrepancy between the meta-training and meta-testing stages. Precisely, we follow the same procedure as (Tseng et al., 2020) where we first train on the whole *mini*-ImageNet dataset using the same setting as detailed above. Then, we evaluate the embeddings model on four different domains: CUB (Welinder et al., 2010), Cars (Krause et al., 2013), Places (Zhou et al., 2017), and Plantae (Van Horn et al., 2018). We show the obtained results in Table 9, and see a notable gain in performance using the proposed method, from 2% gain on CUB dataset, up to 7% gain on Cars dataset, indicating a clear enhancement in terms of the generalization of the embedding model.

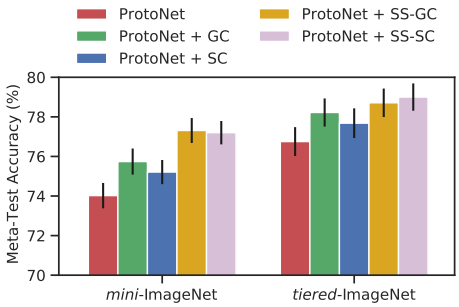

Figure 8. The obtained improvement when adding the contrastive objectives as auxiliary losses. We show the mean acc. and 95% confidence interval for 5-way 5-shot classification across ImageNet derivatives.

Method	Image Size	Backbone	Aux. Loss	Acc. (%)
MAML	84×84	Conv4-64	-	63.1
ProtoNet	84×84	Conv4-64	-	68.2
RelationNet	84×84	Conv4-64	-	65.3
ProtoNet (Chen et al., 2019)	84×84	Conv4-64	-	64.2
ProtoNet (Chen et al., 2019)	224×224	ResNet-18	-	73.7
ProtoNet (Gidaris et al., 2019)	84×84	Conv4-64	-	70.0
		Conv4-64	Rotation	71.7
		Conv4-512	-	71.6
		Conv4-512	Rotation	74.0
		WRN-28-10	-	68.7
		WRN-28-10	Rotation	72.1
ProtoNet (Su et al., 2020)	224×224	ResNet-18	-	75.2
			Rotation	76.0
			Jigsaw	76.2
			Rot.+Jig.	76.6
Ours	224×224	ResNet-18	-	74.0
			GC	75.2
			SC	75.2
			SS-GC	77.3
			SS-SC	77.2
			SS-GC+SS-SC	**77.6**

Table 10. Comparison with prior works on *mini*-ImageNet. We report the mean acc. for 5-shot 5-way classification with implementation details including image size, backbone model and auxiliary training losses for each method.

5. ProtoNet Experiments

To demonstrate the generality of the proposed approach and its applicability in different settings, in this section, we provide additional metric-learning based experiments in which we integrate the contrastive losses into the ProtoNet (Snell et al., 2017) framework. ProtoNet is a distance-based learner trained in an episodic manner, so that both the meta-training and meta-testing stages have matching conditions. During meta-training, for a C-way K-shot setting, we construct a meta-training set $\mathcal{T} = \{(\mathcal{D}_t^{\text{train}}, \mathcal{D}_t^{\text{test}})\}_{t=1}^{T}$ where each given task $(\mathcal{D}_t^{\text{train}}, \mathcal{D}_t^{\text{test}})$ depicts C randomly chosen classes from the seen classes, with K images per class for the training (*i.e.*, support) set $\mathcal{D}_t^{\text{train}}$, and M images per classes for the test (*i.e.*, query) set $\mathcal{D}_t^{\text{test}}$. At each training iteration, after sampling a given task from $\mathcal{T}$, we first compute the class prototypes for classification using the support set. Then, the embeddings model is trained to minimize the CE loss where each query example is classified based on the distances to the class prototypes. In order to add the contrastive objectives as auxiliary losses to the ProtoNet training objective, we simply merge the query and support set, augment each exampled within it, and compute the contrastive losses detailed in §3 over this merged and augmented set.

Experimental Details. For the experimentation, we follow (Chen et al., 2019) and base our implementation on their few-shot learning code base. In particular, we use a ResNet-18 network as the embedding model with 512-dimensional output features. We train on ImagetNet derivatives using ADAM optimizer with a learning rate of 10^{-3} for 60,000 episodes and use 5-way (classes) 5-shot (examples per-class) with 16 query images. For contrastive learning, similar to §4, we use and a two-layer MLP for the projection head and the attention modules with an output dimensionality of 64, and set λ_{CE} to 1.0, and λ_{GC} and λ_{SC} to 0.5. For meta-testing, we report the mean accuracy and 95% confidence interval over 600 randomly sampled tasks, where each class consists of 5 support images and 16 query images.

Results. To investigate the impact of the contrastive losses on the performances of ProtoNet, we report the mean acc. for 5-way 5-shot classification on ImageNet derivatives with different training objectives. The results in Fig. 8 show a notable performance gain over the ProtoNet baseline. Surprisingly, when disregarding the labels and training with the self-supervised formulation of the contrastive objectives, we obtain better results. The SS-SC and SS-GC losses perform comparatively on *mini*-ImageNet with a 3.2% gain, and with the SS-SC loss performing slightly better on *tiered*-ImageNet with a 2.2% gain. We suspect that the obtained gain when using the self-supervised formulation might be a result of using a larger number of negatives as opposed to the supervised formulation, since each batch of examples only contains 5 unique classes. Additionally, we compare the performances of our approach with other self-supervised auxiliary losses, *i.e.*, rotation prediction (Gidaris et al., 2018) and jigsaw puzzle (Noroozi & Favaro, 2016), for which (Su et al., 2020) provided their integration into the ProtoNet framework. As shown in Table 10, we observe that a larger performance gain can be obtained with the contrastive objectives as auxiliary losses compared to other self-supervised objectives, especially when using both the SS-SC and SS-GC losses with a 3.6% gain over the baseline, which further confirms the effectiveness of the proposed SC loss.

6. Related Work

Few-Shot Classification. In few-shot classification, the objective is to learn to recognize unseen novel classes with few labeled example in each class. Meta-learning remains the most popular paradigm to tackle this problem.

Roughly, such approaches can be divided into two categories. Optimization-based, or *learning to learn* methods (Ravi & Larochelle, 2017; Andrychowicz et al., 2016; Wang & Hebert, 2016; Finn et al., 2017; Sun et al., 2019; Lee et al., 2019; Rusu et al., 2019), that integrate the fine-tuning process in the meta-training algorithm to rapidly adapt to model to the unseen classes with limited supervision. And metric-based, or *learning to compare* methods (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018; Oreshkin et al., 2018; Scott et al., 2018; Doersch et al., 2020), that learn a common embedding space in which the similarities between the data can help distinguish between different novel categories with a given distance metric. Most relevant to our work are the methods that follow the standard transfer learning strategy (Chen et al., 2019; Dhillon et al., 2020; Tian et al., 2020b; Afrasiyabi et al., 2020). Consisting of two stages, a pre-training stage with the CE loss on the meta-training set, then a fine-tuning stage on the meta-testing set. Despite their apparent simplicity, (Tian et al., 2020b) showed that such a strategy yields state-of-the-art results on standard benchmarks.

Cross-Entropy Loss. The CE loss continues to be the prominent supervised learning objective used for training deep networks, in which the model is trained to predict the corresponding class label in the form of a one-hot vector. However, despite its success, some works showed many possible drawbacks (Khosla et al., 2020), *e.g.*, noise sensitivity (Sukhbaatar et al., 2014; Zhang & Sabuncu, 2018), adversarial examples (Nar et al., 2019), and suboptimal margins (Elsayed et al., 2018; Cao et al., 2019). While other works proposed some alternative approaches, such as changing the label distribution (Szegedy et al., 2016; Yun et al., 2019; Müller et al., 2019; Zhang et al., 2018) or leveraging the contrastive losses (Khosla et al., 2020).

Contrastive Learning. Instead of training the network to match to the input to a fixed target. Contrastive learning acts directly on the low-dimensional representations with contrastive losses (Hadsell et al., 2006; Gutmann & Hyvärinen, 2010; Salakhutdinov & Hinton, 2007), that measure the similarities of different samples in the embedding space. Recently, contrastive learning based methods have emerged as the state-of-the-art approaches for self-supervised representation learning. The main difference between them is the way they construct and choose the positive samples. In this work, we differentiate between self-supervised contrastive methods (Oord et al., 2018; Wu et al., 2018b; Hjelm et al., 2018; Henaff, 2020; Tian et al., 2019a; He et al., 2020; Chen et al., 2020) that leverage data augmentations to construct the positive pairs, and supervised contrastive methods (Salakhutdinov & Hinton, 2007; Wu et al., 2018a; Kamnitsas et al., 2018; Khosla et al., 2020) that leverage the provided labels to sample the positive examples.

Self-Supervised Learning and Few-Shot Classification. Relevant to our work are methods that try to build on the insights and advances in contrastive learning, or more broadly self-supervised learning, to improve the few-shot classification task. Such methods (Gidaris et al., 2019; Medina et al., 2020; Su et al., 2020; Doersch et al., 2020; Gao et al., 2021) integrate various types of self-supervised training objectives into different few-shot learning frameworks in order to learn more transferable features and improve the few-shot classification performance. In this paper, we propose a novel contrastive learning objective based on the spatial features to further promote general purpose and robust representations suited for few-shot classification. In this context, a similar idea was proposed in (Doersch et al., 2020). In their approach, a constrastive pre-training stage is first conducted followed by the standard ProtoNet (Snell et al., 2017) fine-tuning stage where spatial features are used to compute the similarity between the training and testing instances. In our work, contrary to (Doersch et al., 2020), we integrate the spatial information directly into the contrastive learning loss. The proposed loss is then integrated into the training as an auxiliary loss, resulting in a far more effective, flexible and general framework usable in various few-shot learning scenarios.

Conclusion

In this paper, we investigated contrastive losses as auxiliary training objectives along the CE loss to compensate for its drawbacks and learn richer and more transferable features. With extensive experiments, we showed that integrating contrastive learning into existing few-shot learning frameworks results in a notable boost in performances, especially with our spatial contrastive learning objective. Future work could investigate the spatial contrastive method extension for other few-shot learning scenarios and adapt it for other visual tasks such as unsupervised representation learning.

Acknowledgements

The first author is supported by Randstad corporate research chair, in collaboration with CentraleSupélec, Université Paris-Saclay. This work was performed using HPC resources from the Mésocentre computing center of CentraleSupélec and École NormaleSupérieure Paris-Saclay supported by CNRS and Région Île-de-France (http://mesocentre.centralesupelec.fr/).

References

Afrasiyabi, A., Lalonde, J.-F., and Gagné, C. Associative alignment for few-shot image classification. In *European Conference on Computer Vision*, pp. 18–35. Springer, 2020.

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., Shillingford, B., and De Freitas, N. Learning to

learn by gradient descent by gradient descent. In *Advances in neural information processing systems*, pp. 3981–3989, 2016.

Bertinetto, L., Henriques, J. F., Torr, P. H., and Vedaldi, A. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pp. 1567–1578, 2019.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F., and Huang, J.-B. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Davies, D. L. and Bouldin, D. W. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

Dhillon, G. S., Chaudhari, P., Ravichandran, A., and Soatto, S. A baseline for few-shot image classification. In *International Conference on Learning Representations*, 2020.

Doersch, C., Gupta, A., and Zisserman, A. Crosstransformers: spatially-aware few-shot transfer. In *Advances in Neural Information Processing Systems*, 2020.

Dvornik, N., Schmid, C., and Mairal, J. Diversity with cooperation: Ensemble methods for few-shot classification. In *IEEE International Conference on Computer Vision*, 2019.

Elsayed, G., Krishnan, D., Mobahi, H., Regan, K., and Bengio, S. Large margin deep networks for classification. In *Advances in neural information processing systems*, pp. 842–852, 2018.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 06–11 Aug 2017.

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., and Anandkumar, A. Born again neural networks. 2018.

Gao, Y., Fei, N., Liu, G., Lu, Z., Xiang, T., and Huang, S. Contrastive prototype learning with augmented embeddings for few-shot learning. *arXiv preprint arXiv:2101.09499*, 2021.

Garcia, V. and Bruna, J. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 2018.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. 2018.

Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., and Cord, M. Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8059–8068, 2019.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010.

Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Henaff, O. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4182–4192. PMLR, 13–18 Jul 2020.

Hilliard, N., Phillips, L., Howland, S., Yankov, A., Corley, C. D., and Hodas, N. O. Few-shot learning with metric-agnostic conditional embeddings. *arXiv preprint arXiv:1802.04376*, 2018.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018.

Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017.

Kamnitsas, K., Castro, D. C., Folgoc, L. L., Walker, I., Tanno, R., Rueckert, D., Glocker, B., Criminisi, A., and Nori, A. Semi-supervised learning via compact latent space clustering. *arXiv preprint arXiv:1806.02679*, 2018.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, 2020.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research). *URL http://www. cs. toronto. edu/kriz/cifar. html*, 5, 2010.

Lake, B., Salakhutdinov, R., Gross, J., and Tenenbaum, J. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

Lee, K., Maji, S., Ravichandran, A., and Soatto, S. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10657–10665, 2019.

Li, H., Dong, W., Mei, X., Ma, C., Huang, F., and Hu, B.-G. LGM-net: Learning to generate matching networks for few-shot learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3825–3834. PMLR, 09–15 Jun 2019.

Li, Z., Zhou, F., Chen, F., and Li, H. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

Medina, C., Devos, A., and Grossglauser, M. Self-supervised prototypical transfer learning for few-shot classification. *arXiv preprint arXiv:2006.11325*, 2020.

Mishra, N., Rohaninejad, M., Chen, X., and Abbeel, P. A simple neural attentive meta-learner. In *International Conference on Learning Representations*, 2018.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pp. 4694–4703, 2019.

Naik, D. K. and Mammone, R. J. Meta-neural networks that learn by learning. In *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*, volume 1, pp. 437–442. IEEE, 1992.

Nar, K., Ocal, O., Sastry, S. S., and Ramchandran, K. Cross-entropy loss and low-rank features have responsibility for adversarial examples. *arXiv preprint arXiv:1901.08360*, 2019.

Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pp. 69–84. Springer, 2016.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Oreshkin, B., López, P. R., and Lacoste, A. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 721–731, 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Qi, H., Brown, M., and Lowe, D. G. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5822–5830, 2018.

Qiao, L., Shi, Y., Li, J., Wang, Y., Huang, T., and Tian, Y. Transductive episodic-wise adaptive metric for few-shot learning. In *IEEE International Conference on Computer Vision*, 2019.

Raghu, A., Raghu, M., Bengio, S., and Vinyals, O. Rapid learning or feature reuse? towards understanding the effectiveness of maml. In *International Conference on Learning Representations*, 2019.

Ravi, S. and Larochelle, H. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2017.

Ravichandran, A., Bhotika, R., and Soatto, S. Few-shot learning with embedded class models and shot-free meta training. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 331–339, 2019.

Ren, M., Ravi, S., Triantafillou, E., Snell, J., Swersky, K., Tenenbaum, J. B., Larochelle, H., and Zemel, R. S. Meta-learning for semi-supervised few-shot classification. In *International Conference on Learning Representations*, 2018.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Rusu, A. A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., and Hadsell, R. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019.

Salakhutdinov, R. and Hinton, G. Learning a nonlinear embedding by preserving class neighbourhood structure. In *Artificial Intelligence and Statistics*, pp. 412–419, 2007.

Scott, T., Ridgeway, K., and Mozer, M. C. Adapted deep embeddings: A synthesis of methods for k-shot inductive transfer learning. In *Advances in Neural Information Processing Systems*, pp. 76–85, 2018.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pp. 4077–4087, 2017.

Su, J.-C., Maji, S., and Hariharan, B. When does self-supervision improve few-shot learning? In *European Conference on Computer Vision*, pp. 645–666. Springer, 2020.

Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., and Fergus, R. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

Sun, Q., Liu, Y., Chua, T.-S., and Schiele, B. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 403–412, 2019.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Thrun, S. Lifelong learning algorithms. In *Learning to learn*, pp. 181–209. Springer, 1998.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *International Conference on Learning Representations*, 2019a.

Tian, Y., Krishnan, D., and Isola, P. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019b.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning. In *Advances in Neural Information Processing Systems*, 2020a.

Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J. B., and Isola, P. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*. Springer, 2020b.

Tseng, H.-Y., Lee, H.-Y., Huang, J.-B., and Yang, M.-H. Cross-domain few-shot classification via learned feature-wise transformation. In *International Conference on Learning Representations*, 2020.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Wang, Y.-X. and Hebert, M. Learning to learn: Model regression networks for easy small sample learning. In *European Conference on Computer Vision*, pp. 616–634. Springer, 2016.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-ucsd birds 200. 2010.

Wu, Z., Efros, A. A., and Yu, S. X. Improving generalization via scalable neighborhood component analysis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 685–701, 2018a.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018b.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6023–6032, 2019.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*, pp. 8778–8788, 2018.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Supplementary Material

A. Datasets

In this paper, we experimented with different datasets used for various few-shot classification settings. For the standard few-shot classification setting, we used the four popular benchmarks: *mini*-ImageNet, *tiered*-ImageNet, CIFAR-FS and FC100. As for cross-domain few-shot classification, following (Tseng et al., 2020), we train on the whole *mini*-ImageNet (*i.e.*, train, val and test sets), and then we evaluate on one of the following datasets: CUB, Cars, Places or Plantae. The details about each dataset are presented in Table A.1.

Datasets	Source	Nbr. of train classes	Nbr. of val classes	Nbr. of test classes	Split setting
mini-ImageNet	(Vinyals et al., 2016)	64	16	20	(Ravi & Larochelle, 2017)
tiered-ImageNet	(Ren et al., 2018)	351	97	160	Original
CIFAR-FS	(Bertinetto et al., 2019)	64	16	20	Original
FC100	(Oreshkin et al., 2018)	60	20	20	Original
CUB	(Welinder et al., 2010)	100	50	50	(Hilliard et al., 2018)
Cars	(Krause et al., 2013)	98	49	49	Random
Places	(Zhou et al., 2017)	183	91	91	Random
Plantae	(Van Horn et al., 2018)	100	50	50	Random

Table A.1. Datasets. Additional details about the datasets used in the experiments.

B. Additional Experiments

B.1. Embedding Model

In the transfer learning experiments, we mainly used a ResNet-12 backbone as the embedding model. Since the backbone also has a significant impact on the quality of the produced embedding, we experiment with various backbones. In particular, we compare ResNet-12 to two other alternative: 4 layer convolution network (*i.e.*, 64-64-64-64) and ResNet-12 with sequeeze-and-excitation (Hu et al., 2018) layers that play the role of a channel-wise attention module. See Fig. D.1 for an illustration of these backbones. The results are presented in Table B.1, where we train the models either on the meta-training set only or on both the meta-training and the meta-validation set. In both cases with and without a contrastive distillation step. As expected, we observe that, (1) better models improve the performances, and (2) more training data yield better results, further emphasizing the importance of learning a well performing and transferable embedding model for an effective few-shot classification.

B.2. Weight Imprinting

During the ablation experiments, and when training a base classifier on top of the spatial features, a case that requires a larger classifier, we observed a slight decrease in the performances, which we suspect might be due to the overfitting of the model, especially at low shot setting. To overcome this, we investigate using imprinted weights (Qi et al., 2018) to initialize the base classifier and help stabilize the convergence. Weight imprinting consists of directly setting the base classifier's weights from novel training examples during low-shot learning *i.e.*, the ℓ_2 normalized class prototypes constructed from the training (*i.e.*, support) features. This process is called weight imprinting since it directly sets the weights of the base classifier based on an the scaled features produced by the embedding model for a given training example at test time. Table B.2 summarizes the results, although the weight imprinting does slightly help the performances, the results are not conclusive and need further investigation.

B.3. Contrastive Distillation

As detailed in the paper, in order to relax the placement of the embeddings when applying a distillation step, we also optimize a contrastive loss L_{CD} loss in addition to the standard KL loss. The loss consists of aligning the global features of the teacher and the student by reducing their cosine similarity. However, we can also align the spatial features rather than the global features, or both. In this case, the contrastive loss is written as follows:

$$L_{\text{CD}} = \alpha \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{z}_i^{\text{g}t} - \mathbf{z}_i^{\text{g}s}\|_2^2 + \beta \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{z}_i^{\text{s}t} - \mathbf{z}_i^{\text{s}s}\|_2^2. \tag{12}$$

The results are shown in Table B.3. We observe while contrasting only the global features works slightly better, the obtained

Method	Backbone	*mini*-ImageNet, 5-way		CIFAR-FS, 5-way		FC100, 5-way	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Ours	64-64-64-64	50.45 ± 0.83	67.88 ± 0.61	60.31 ± 0.92	77.15 ± 0.66	38.45 ± 0.72	52.15 ± 0.77
Ours-Distill	64-64-64-64	52.18 ± 0.89	68.45 ± 0.68	60.8 ± 0.88	77.25 ± 0.67	38.31 ± 0.75	52.06 ± 0.77
Ours-Trainval	64-64-64-64	51.74 ± 0.76	68.69 ± 0.74	60.75 ± 0.92	77.56 ± 0.72	$\mathbf{42.07 \pm 0.74}$	55.97 ± 0.73
Ours-Distill-Trainval	64-64-64-64	$\mathbf{52.64 \pm 0.84}$	$\mathbf{69.46 \pm 0.61}$	$\mathbf{62.02 \pm 0.93}$	$\mathbf{77.88 \pm 0.76}$	41.73 ± 0.75	$\mathbf{56.16 \pm 0.75}$
Ours	ResNet-12	65.66 ± 0.76	82.52 ± 0.50	75.01 ± 0.91	87.44 ± 0.58	44.30 ± 0.70	59.80 ± 0.70
Ours-Distill	ResNet-12	67.40 ± 0.76	82.70 ± 0.52	76.46 ± 0.87	87.62 ± 0.59	44.80 ± 0.70	61.40 ± 0.70
Ours-Trainval	ResNet-12	67.02 ± 0.81	84.01 ± 0.53	76.08 ± 0.87	87.69 ± 0.62	50.89 ± 0.79	67.57 ± 0.75
Ours-Distill-Trainval	ResNet-12	$\mathbf{68.54 \pm 0.84}$	$\mathbf{84.50 \pm 0.50}$	$\mathbf{77.88 \pm 0.81}$	$\mathbf{88.57 \pm 0.61}$	$\mathbf{51.12 \pm 0.81}$	$\mathbf{67.96 \pm 0.69}$
Ours	SEResNet-12	65.83 ± 0.76	81.66 ± 0.58	74.70 ± 0.92	86.90 ± 0.59	42.45 ± 0.71	59.72 ± 0.71
Ours-Distill	SEResNet-12	66.38 ± 0.81	83.25 ± 0.50	76.41 ± 0.88	87.44 ± 0.62	43.00 ± 0.73	60.84 ± 0.76
Ours-Trainval	SEResNet-12	67.84 ± 0.72	83.38 ± 0.55	75.54 ± 0.87	87.27 ± 0.60	$\mathbf{51.70 \pm 0.78}$	68.19 ± 0.73
Ours-Distill-Trainval	SEResNet-12	$\mathbf{69.28 \pm 0.84}$	$\mathbf{83.92 \pm 0.52}$	$\mathbf{76.72 \pm 0.82}$	$\mathbf{87.86 \pm 0.59}$	51.06 ± 0.83	$\mathbf{68.28 \pm 0.73}$

Table B.1. Comparison of different backbones on 5-way classification on *mini*-ImageNet, CIFAR-FS and FC100. We show the mean acc. and 95% confidence interval.

Features Used	Weight Imp.	*mini*-ImageNet, 5-way		CIFAR-FS, 5-way		FC100, 5-way	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Spatial		63.40 ± 0.86	81.42 ± 0.52	74.56 ± 0.89	86.83 ± 0.64	42.92 ± 0.71	59.49 ± 0.74
Global		64.49 ± 0.85	81.87 ± 0.58	74.96 ± 0.89	87.65 ± 0.59	43.90 ± 0.77	60.61 ± 0.76
Glo. & Spa. (Max)		64.94 ± 0.78	82.18 ± 0.50	74.40 ± 0.87	87.18 ± 0.58	$\mathbf{43.56 \pm 0.75}$	$\mathbf{60.78 \pm 0.71}$
Glo. & Spa. (Sum)		65.30 ± 0.77	81.53 ± 0.55	74.30 ± 0.85	$\mathbf{87.72 \pm 0.58}$	43.18 ± 0.73	60.11 ± 0.70
Spatial	✓	64.12 ± 0.82	81.98 ± 0.53	74.93 ± 0.81	87.30 ± 0.59	43.87 ± 0.71	60.64 ± 0.82
Global	✓	65.14 ± 0.81	$\mathbf{82.72 \pm 0.58}$	$\mathbf{75.43 \pm 0.88}$	87.18 ± 0.56	43.21 ± 0.77	59.53 ± 0.72
Glo. & Spa. (Max)	✓	64.91 ± 0.79	81.60 ± 0.57	74.39 ± 0.84	87.22 ± 0.60	43.54 ± 0.72	59.44 ± 0.77
Glo. & Spa. (Sum)	✓	$\mathbf{65.56 \pm 0.84}$	82.43 ± 0.54	75.08 ± 0.87	87.30 ± 0.57	$\mathbf{43.56 \pm 0.68}$	60.04 ± 0.70

Table B.2. Comparison of different evaluation setups. We train a linear classifier on either the spatial, the global features, or both, where two classifiers are used and their predictions are aggregated by either taking their sum or maximum per class. We also investigate the effect of weight imprinting to improve the initialization of the classifier at test time. Note that the presented results use a PyTorch implementation of the classifiers instead of scikit-learn, resulting in slight decrease in performances.

Contrastive Distillation	*mini*-ImageNet, 5-way		CIFAR-FS, 5-way		FC100, 5-way	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Teacher	*65.7 ± 0.8*	*82.5 ± 0.5*	*75.0 ± 0.9*	*87.4 ± 0.6*	*44.4 ± 0.8*	*60.8 ± 0.9*
Global ($\beta = 0$)	$\mathbf{67.4 \pm 0.8}$	$\mathbf{83.2 \pm 0.5}$	$\mathbf{76.5 \pm 0.9}$	$\mathbf{88.0 \pm 0.6}$	$\mathbf{44.8 \pm 0.7}$	$\mathbf{61.4 \pm 0.7}$
Spatial ($\alpha = 0$)	66.0 ± 0.8	$\mathbf{83.2 \pm 0.5}$	76.0 ± 0.8	$\mathbf{80.0 \pm 0.5}$	44.4 ± 0.7	61.0 ± 0.7
Global & Spatial	65.6 ± 0.8	83.1 ± 0.5	75.4 ± 0.9	87.3 ± 0.6	44.3 ± 0.8	61.2 ± 0.8

Table B.3. Comparison of contrastive distillation losses on 5-way classification on *mini*-ImageNet, CIFAR-FS and FC100. We show the mean acc. and 95% confidence interval.

results are overall similar.

C. Additional Experimental Details

Training. As stated in the paper, during training, we reduce the learning rate with a factor of 0.1 at different training iterations. For *mini*-ImageNet, we train for 90 epochs and we decay the learning rate at 60 and 80 epochs, for *tiered*-ImageNet, we train for 60 epochs and decay the learning rate rate three times, at 30, 40 and 50 epochs. As for CIFAR-100 derivatives, we train for 90 epochs and decay the learning rate three times at 45, 60 and 75 epochs for CIFAR-FS, while for FC-100, we train for 65 epochs with single learning decay step at 60 epochs. For distillation, we change to learning rate to 10^{-2} and train with similar settings.

For the weights of the loss functions, while the weight of the cross-entropy loss is always set to 1, when training with only one contrastive loss as an auxiliary loss, we set its weight to 1, *i.e.*, $\lambda_{\mathrm{GC}} = \lambda_{\mathrm{SC}} = 1$, be it the supervised of self-supervised formulations, except for CIFAR-FS where we set $\lambda_{\mathrm{GC}} = \lambda_{\mathrm{SC}} = 0.5$. Additionally, when training with both objectives, we set $\lambda_{\mathrm{SC}} = \lambda_{\mathrm{GC}} = 0.5$.

For transfer learning experiments, we train on either 84×84 sized images for ImageNet derivatives, resulting in spatial features of spatial dimensions of 5×5, or 32×32 sized images for CIFAR-100 derivatives, resulting in spatial features of spatial dimensions of 2×2. For ProtoNet experiments, we train on 224×224 sized images with ResNet-18 backbone,

resulting in spatial features of spatial dimensions of 7×7. Thus, in order to reduce the computational requirement when applying the attention based spatial alignment, we apply an adaptive average pooling, reducing the dimensions to 3×3 instead of 7×7 resulting in both better results and faster training time.

Evaluation. During evaluation, we use a multivariate logistic regression implemented in scikit-learn (Pedregosa et al., 2011) trained on the ℓ_2 normalized global features. Additionally, when using the spatial features during evaluation, and in order to reduce the number of parameters of the linear classifier and avoid overfitting, we first apply a max pooling operation over these features, reducing their spatial dimensions to 2×2, and then feed them to the linear classifier. As for cross-domain experiments, we found that augmenting each training (*i.e.*, support) samples 10 times instead of 5 produces slightly better results.

When conducting ablation studies with weight imprinting (*i.e.*, Table B.2), we use our own implementation of the multivariate logistic regression with an L-BFGS optimizer and an ℓ_2 penalty. While the performances are overall similar, this gives us more degrees of freedom when implementing the base classifier. Specifically, when using the spatial features as input to the base classifier, we implement the classifier as a convolutional layer, in which the filter size matches the dimensions of the spatial features.

D. Architectures

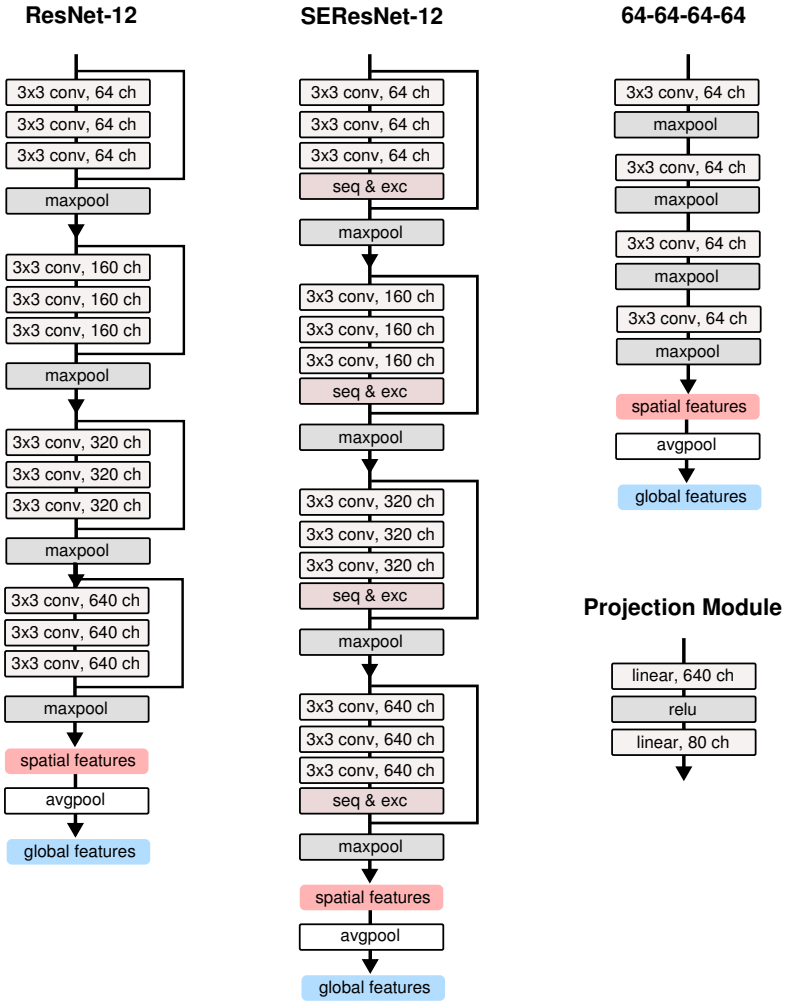

Figure D.1. Architectural details. The architecture of the backbones used as embeddings models, and the projection module used both as the projection head for GC loss, and for the attention modules (*i.e.*, head, query and value heads) for the SC loss. "seq & exc" refers to squeeze and excitation modules.

E. Quality of the Learned Representations

In this section, and similar to the analysis conducted in the paper, we conduct an empirical analysis of the embeddings to assess the quality of the learned features in two cases: (1) when the model is trained with only the CE loss, and (2) when adding the SC as an additional auxiliary objective. Figs. E.1 and E.2 show the results. We observe a clear improvement in terms of the obtained nearest neighbors when using the SC loss and also an increase in terms of the amount of informative signal retrained within the embedding matrix, both indicating an enhancement in the quality of the learned embeddings.

Figure E.1. Nearest Neighbors Analysis. For a given test image from *mini*-ImageNet dataset, we compute the nearest neighbors in the embedding space on the test set when a model is train with either, (a) the standard CE, or, (b) with the proposed SC as an additional auxiliary objective. We observe that the neighboring images in the embedding space found when the SC loss is used are more semantically similar then the standard case with the CE loss. It suggests that the quality of the learned embeddings is increased with the usage of the SC as an auxiliary loss as a result of optimizing for more general-purpose features.

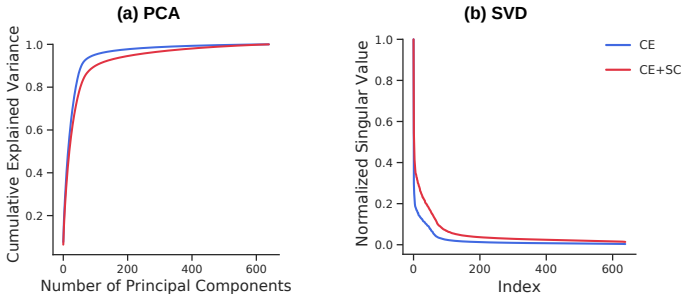

Figure E.2. Spectral Analysis. Results of the spectral analysis on the embedding matrix using CE or CE+SC as training objectives. The plot (a) shows the explained cumulative variance of the learned features as the number of principal components used and (b) shows the max-normalized singular values. We observe that the SC loss increases the number of dominant principal components and the weight of the remaining singular values, which indicates that the SC loss does help retain more informative signals that might be useful outside of the meta-training classification task.