DMN4: Few-Shot Learning via Discriminative Mutual Nearest Neighbor Neural Network

Yang Liu[1], **Tu Zheng**[1,2], **Jie Song**[3], **Deng Cai**[1,2], **Xiaofei He**[1,2]
[1]State Key Lab of CAD&CG, College of Computer Science, Zhejiang University
[2]Fabu Inc., Hangzhou, China
[3]Zhejiang University
{lyng_95, sjie}@zju.edu.cn, zhengtuzju@gmail.com, dengcai@cad.zju.edu.cn, xiaofeihe@fabu.ai

Abstract

Few-shot learning (FSL) aims to classify images under low-data regimes, where the conventional pooled global feature is likely to lose useful local characteristics. Recent work has achieved promising performances by using deep descriptors. They generally take all deep descriptors from neural networks into consideration while ignoring that some of them are useless in classification due to their limited receptive field, *e.g.*, task-irrelevant descriptors could be misleading and multiple aggregative descriptors from background clutter could even overwhelm the object's presence. In this paper, we argue that a Mutual Nearest Neighbor (MNN) relation should be established to explicitly select the query descriptors that are most relevant to each task and discard less relevant ones from aggregative clutters in FSL. Specifically, we propose Discriminative Mutual Nearest Neighbor Neural Network (DMN4) for FSL. Extensive experiments demonstrate that our method outperforms the existing state-of-the-arts on both fine-grained and generalized datasets.

Introduction

With the availability of large-scale training data, deep neural networks have achieved great success in recent years (He et al. 2016; Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2015). However, collecting and labeling training data are still laboriously painful. In terms of low-data scenarios, such as medical images and endangered species, deep neural networks can easily collapse. Few-shot learning (FSL), whose goal is to construct a model that can be readily adapted to novel classes given just a small number of labeled instances, has emerged as a promising paradigm to alleviate this problem.

Few-shot learning methods can be roughly categorized into two schools, *i.e.*, optimization based (Finn, Abbeel, and Levine 2017; Rusu et al. 2018) and metric learning based (Vinyals et al. 2016; Snell, Swersky, and Zemel 2017; Sung et al. 2018; Li et al. 2019). Optimization based methods aim to learn a good parameter initialization for the classifier, whose weights can be quickly adapted to novel classes using gradient-based optimization on only a few labeled samples. Metric learning methods mainly focus on concept representation or relation measures by learning a deep embedding space to transfer knowledge. They generally treat deep pooled features from the global average pooling layer as an image-level representation, which is a common practice for large-scale image classification. Considering the unique characteristic of FSL (*i.e.*, the scarcity of examples for each class), however, the cluttered background and large intra-class variations would drive these pooled global features from the same category far apart in a given metric space under low-data regimes, where useful local characteristics could be overwhelmed and lost.

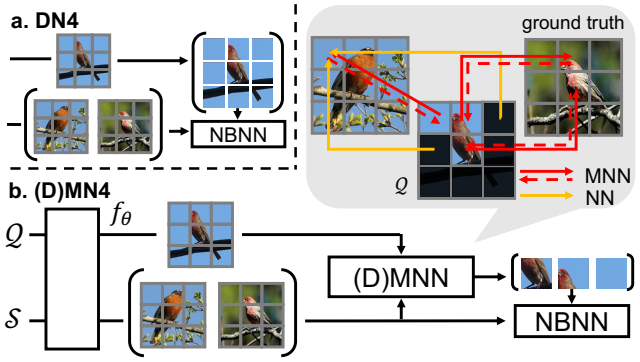


Figure 1: DN4 (Li et al. 2019) accumulates all query descriptors where multiple "*sky*" descriptors are taken as strong evidence against *birds*' presence. (D)MN4 selects discriminative task-relevant query descriptors (grids connected by double-ended red lines) by introducing MNN where less relevant query descriptors (the shaded grids) nearest neighboring to the same support descriptor would be ignored if they were not the mutual nearest one.

To fully exploit the local characteristics, Li et al. revisited Naive-Bayes Nearest Neighbor (NBNN) (Boiman, Shechtman, and Irani 2008) to retain all reference deep descriptors in their original form. They remove the last global average pooling layer to achieve a dense image representation and treat the output feature map as a set of deep local descriptors. For each descriptor from a query image, they calculate its similarity scores to the nearest neighbor descriptors in each support class. Finally, similarity scores from all query descriptors are accumulated as an image-to-class similarity.

However, in our perspective, there is a notable differ-

ence between local invariant descriptors (*e.g.*, SIFT) in traditional NBNN and network deep descriptors: the former one is position-agnostic and diversely distributed (around salient positions), while deep descriptors from neural networks are densely distributed like a grid. Directly accumulating all deep descriptors violates the intuition that the presence of background clutter shouldn't be taken as a strong evidence against the object's presence.

Although the background clutters influence the NBNN classification, there is rarely a straightforward way to preselect those backgrounds unless introducing extra modules for the foreground retrieval. Instead, we try to mitigate the influences from those descriptors in a different way by recognizing the fact that descriptors within a background clutter are relatively similar to their nearby descriptors, *e.g.*multiple local characterless *blue sky* in Figure 1 themselves are quite similar compared to the huge difference between the characteristic *beak* and *wings* of the *bird*.

Based on this observation, we introduce Mutual Nearest Neighbor (MNN) for NBNN in this paper to mitigate the accumulated influences from aggregative background clutters so that less relevant characterless background descriptors account less during classification. To further mine the discriminative descriptors in classification, we propose a novel Discriminative Mutual Nearest Neighbor (DMNN) algorithm based on the derivation of NBNN, which is quantitatively shown to be effective in the experiments. In summary, the contributions are: (1) We propose to find discriminative descriptors to improve NBNN based few-shot classification. To the best of our knowledge, this is the first attempt to combine MNN with NBNN in deep learning framework. (2) The proposed methods outperform the state-of-the-art on both fine-grained and generalized few-shot classification datasets. The proposed methods could be easily extended to a semi-supervised version without extra bells or whistles.

Related Work

Global Feature based methods. Traditional metric learning based methods use compact feature vectors from the last global average pooling layer of the network to represent images and classification is performed via simple classifiers or nearest neighbors directly. Vinyals et al. trains a learnable nearest neighbor classifier with a deep neural network. Snell, Swersky, and Zemel takes the mean of each class as its corresponding prototype representation to learn a metric space. Sung et al. introduces an auxiliary non-linear metric to compute the similarity score between each query and support set. These deep global features would lose considerable discriminative local information under low-data regimes.

Deep Descriptor based methods. Another branch of metric learning methods focuses on using deep descriptors to solve few-shot classification. Lifchitz et al. propose to make predictions for each local representation and average their output probabilities. Huang et al. combines local descriptors with prototypical learning. Zhang et al. adopts the earth mover's distance as a metric to compute a structural distance between dense image representations to determine image relevance. Li et al. uses the top k nearest vectors between two feature maps in a Naive-Bayes way to represent image-level distance. Our DMN4 further highlights the importance of selecting discriminative and task-relevant descriptors in the deep descriptors based method.

Subspace Learning based methods. Recent works also investigate the potential of adaptive subspace learning in FSL. Yoon, Seo, and Moon learns a task-specific subspace projection and the classification is performed based on the mapped query features and projected references. Simon et al. learns class-specific subspaces based on the few examples within each class, and the classification is performed based on the shortest distance among query projections onto each subspace. Both of them adapt subspace projection with few examples provided in each task but ignore that projection matrices derived from matrix decomposition could easily collapse under low-data regimes. DMN4 could also be treated among the family of subspace learning as it also selects a subset of descriptors for each query example. Differently, our subspace dimensionality will be automatically determined by the number of MNN pairs instead of predefining a hyper-parameter as in the previous literature. Also, the large quantity of descriptors makes it more reliable to retain useful local characteristics compared to the matrix decomposition on a global feature vector.

Methodology

In this work, we focus on the N-way K-shot few-shot classification problem, where N is the number of categories with K labeled examples in each. The model is trained with a large training dataset $\mathcal{D}_{train}$ of labeled examples from classes C_{train} with an episodic training mechanism (Vinyals et al. 2016). In each episode, we first construct a support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N\times K}$ and a query set $\mathcal{Q} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{q}$ containing different samples from the same label, where $\mathcal{S}$ and $\mathcal{Q}$ are sampled from $\mathcal{D}_{train}$; then the model is updated on the labeled $\mathcal{S}$ by minimizing classification loss on $\mathcal{Q}$.

Deep Descriptor based Image Representation

We embed the image x via the backbone network to obtain the 3D feature map representation $f(x) \in \mathbb{R}^{h\times w\times d}$, where $f(\cdot)$ is the hypothesis function of the deep backbone network. Like other descriptor based methods, we treat the feature map as $r = h \times w$ number of d-dimensional descriptors.

There are K-shot images for each support class within an episode. When $K > 1$, some methods use the empirical mean of K compact image representations for the stability and memory efficient in meta-training. Others instead unite those $K \times r$ feature vectors from the same support class to retain descriptors in their original form. In this work, we use the empirical mean of descriptors that are from the deeper feature extractor (*e.g.*, ResNet-12) while unit in their original form for those that are from the shallower backbone network (*e.g.*, Conv-4).

Formally, we denote the set of descriptors from the same support class $c \in C$ as $\mathbf{s}_c$ and denote descriptors from each query image as $\mathbf{q}$. We use the bold font $\{\mathbf{q}, \mathbf{s}\}$ to represent a set of descriptors and $\{q, s\}$ to represent a single channel-dimensional descriptor vector in the following sections.

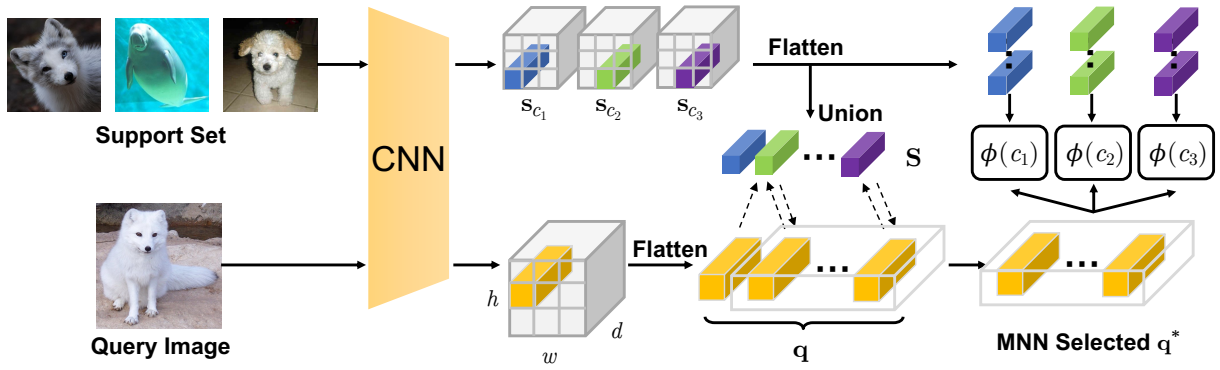


Figure 2: The architecture of **M**utual **N**earest **N**eighbor **N**eural **N**etwork (MN4) for few-shot classification. The episodic data is first fed into embedding CNN to get a deep compact feature map and then flatten as sets of channel-dimensional local descriptors $\mathbf{q}$ and $\mathbf{s}_c$. Support descriptors from different classes unite as a single support pool $\mathbf{S}$. We select part of the query descriptors $\mathbf{q}^*$ by performing MNN between $\mathbf{q}$ and $\mathbf{S}$. The selected $\mathbf{q}^*$ as well as the native $\mathbf{s}_c$ are used to calculate a Naive-Bayes classification score $\phi(c)$ for class c. (The architecture of DMN4 is the same as MN4 except for using DMNN instead of MNN in selectivity.)

Mutual Nearest Neighbor

As discussed, if we directly accumulate all descriptors in a Naive-Bayes way, background clutters and outliers would mislead the classification. To alleviate it, we revisit the concept of Mutual Nearest Neighbor (MNN) (Gowda and Krishna 1979) that initially proposed to obtain a condensed training set decades ago. Formally, we use a single merged support descriptor pool $\mathbf{S} = \bigcup_{c\in C} \mathbf{s}_c$ comprising support descriptors from all classes. For each descriptor $q \in \mathbf{q}$, we find its nearest neighbor $s = \mathrm{NN}_{\mathbf{S}}(q)$ from the support descriptor pool $\mathbf{S}$ and use s to search back its nearest neighbor $\tilde{q} = \mathrm{NN}_{\mathbf{q}}(s)$ from $\mathbf{q}$. If q equals $\tilde{q}$, we consider q and s a MNN pair between query descriptor set $\mathbf{q}$ and support descriptor pool $\mathbf{S}$.

Naive-Bayes Nearest Neighbor for FSL

To help motivate and justify our updates to the original NBNN algorithm, we briefly provide an overview of the original NBNN derivation and its application in DN4 (Li et al. 2019). We start by classifying image x to class c by:

$$\hat{c} = \arg\max_{c\in C} p(c|x) = \arg\max_{c\in C} \log p(c|x). \tag{1}$$

Applying Bayes' rule with the equal class prior and conditional independence assumptions on Eqn.(1) gives:

$$\hat{c} = \arg\max_{c\in C} \left[\log(\prod_{q\in\mathbf{q}} p(q|c))\right] = \arg\max_{c\in C} \left[\sum_{q\in\mathbf{q}} \log p(q|c)\right] \tag{2}$$

We then approximate $p(q|c)$ in Eqn.(2) by a Parzen window estimator with kernel κ:

$$p(q|c) = \frac{1}{|\mathbf{s}_c|}\sum_{j=1}^{|\mathbf{s}_c|} \kappa(q, \mathrm{NN}_{\mathbf{s}_c}(q, j)) \approx \kappa(q, \mathrm{NN}_{\mathbf{s}_c}(q)) \tag{3}$$

where $|\mathbf{s}_c|$ is the cardinality of support descriptor set $\mathbf{s}_c$ and $\mathrm{NN}_{\mathbf{s}_c}(q, j)$ is the j-th nearest descriptor of support class c. NBNN takes it to the extreme by considering only the first nearest neighbor $\mathrm{NN}_{\mathbf{s}_c}(q)$.

Li et al. chooses a *cosine similarity* for the approximation of $\log \kappa(\cdot))$ and substitutes Eqn.(3) into (2) to find the class with the maximum accumulated similarities:

$$\hat{c} = \arg\max_{c\in C} \left[\sum_{q\in\mathbf{q}} \cos(q, \mathrm{NN}_{\mathbf{s}_c}(q))\right] \tag{4}$$

In this work, we further select subspaces $\mathbf{q}^* \in \mathbb{R}^{|\mathbf{q}^*|\times C}$ that owns a relatively stronger bond of mutual closeness with support descriptors. The accumulated similarity score of a query image x to class c in proposed MN4 is

$$\phi(x, c) = \sum_{q\in\mathbf{q}^*} \cos(q, \mathrm{NN}_{\mathbf{s}_c}(q)) \tag{5}$$

and the cross-entropy loss is used to meta-train the network:

$$p(c|x) = \frac{e^{\phi(x,c)}}{\sum_{c'\in C} e^{\phi(x,c')}} \tag{6}$$

$$\mathcal{L} = -\frac{1}{|\mathcal{Q}|}\sum_{\mathcal{Q}}\sum_{c\in C} y \log p(c|x) \tag{7}$$

Towards Discriminative Mutual Nearest Neighbor

MNN selects task-relevant descriptors by considering their mutual *absolute* mutual similarity. Yet, it offers no theoretical guarantee that the selected query descriptors are discriminative enough in NBNN classifications. In this section, we propose a novel *relative closeness* in MNN that designed for NBNN classifications and term it Discriminative Mutual Nearest Neighbor (DMNN). The discriminability indicates how query descriptor q relates to its neighbored support descriptor s than other descriptors in $\mathbf{S}$.

We start by recasting NBNN updates as an adjustment to the posterior log-odds (McCann and Lowe 2012). Let c be some class and $\bar{c}$ be the set of all other classes, the odds ($\mathcal{O}$) for class c is given by:

$$\mathcal{O}_c = \frac{p(c|x)}{p(\bar{c}|x)} = \frac{p(x|c)p(c)}{p(x|\bar{c})p(\bar{c})} = \prod_{q\in\mathbf{q}} \frac{p(q|c)}{p(q|\bar{c})}\frac{p(c)}{p(\bar{c})} \tag{8}$$

This allows an alternative classification rule expressed in terms of log-odds increments and class priors:

$$\hat{c} = \arg\max_{c \in C} \left[\sum_{q \in \mathbf{q}} \log \frac{p(q|c)}{p(q|\bar{c})} + \log \frac{p(c)}{p(\bar{c})} \right] \tag{9}$$

Approximating by a Parzen window estimator like in Eqn.(3) and assuming an equal class prior give the NBNN log-odds classification rule (find the class with the largest accumulated relative similarities):

$$\hat{c} = \arg\max_{c \in C} \sum_{q \in \mathbf{q}} \left(\cos(q, \mathrm{NN}_{\mathbf{s}_c}(q)) - \cos(q, \mathrm{NN}_{\mathbf{S} \setminus \mathbf{s}_c}(q)) \right) \tag{10}$$

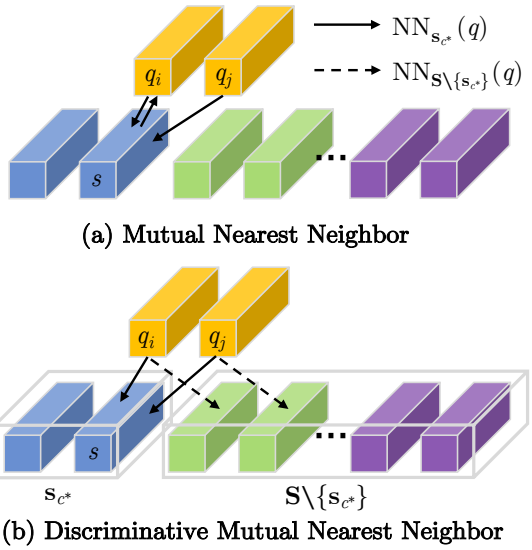
where $\mathbf{S} \setminus \mathbf{s}_c$ represents all support descriptors set minus the descriptors from class c.

Recall that the basic idea of MNN is equivalent to discarding quantities of characterless descriptors. To further guarantee the discriminability of selected descriptors, we take their relative closeness into consideration. Formally, for each query descriptor $q \in \mathbf{q}$, we first find the belonging class c^* of $s \in \mathbf{S}$ that is nearest to q, *i.e.*, $\mathbf{s}_{c^*} \ni s = \mathrm{NN}_{\mathbf{S}}(q)$. To measure whether a query descriptor q discriminative enough in MNN selection, we consider its relative closeness $\tau(q)$ that represents how q votes for its nearest support class c^* than the other supporting classes $C \setminus \{c^*\}$:

$$c^* = \arg\max_{c \in C} \cos(q, \mathrm{NN}_{\mathbf{s}_c}(q)) \tag{11}$$

$$\tau(q) = \cos(q, \mathrm{NN}_{\mathbf{s}_{c^*}}(q)) - \cos(q, \mathrm{NN}_{\mathbf{S} \setminus \mathbf{s}_{c^*}}(q)) \tag{12}$$

As illustrated in Figure 3, if both query descriptors q_i, q_j are nearest neighboring to the same support descriptor s in support descriptor pool $\mathbf{S}$, the selectivity is determined by their relative closeness $\tau(q)$ in DMNN while determined by the absolute closeness $\cos(q, s)$ in MNN.



(a) Mutual Nearest Neighbor

(b) Discriminative Mutual Nearest Neighbor

Figure 3: Comparison between MNN and DMNN when multiple query descriptors nearest neighboring to the same support descriptor. (a) MNN selects q by its *absolute* similarity. (b) DMNN selects q by its largest *relative* similarity.

Experiments

Datasets

***mini*ImageNet** (Vinyals et al. 2016) is a subset of ImageNet containing randomly selected 100 classes. We follow the setup provided by Sachin and Hugo that takes 64, 16 and 20 classes for training, validation and evaluation respectively.

***tiered*Imagenet** is a larger subset of ImageNet but contains a broader set of classes compared to the *mini*ImageNet. There are 351 classes from 20 different categories for training, 97 classes from 6 different categories for validation, and 160 classes from 8 different categories for testing (Ren et al. 2018), where the information overlap between training and validation/testing tasks is minimized.

Caltech-UCSD Birds-200-2011 (CUB) (Wah et al. 2011) is a fine-grained dataset that contains 11788 images of 200 birds species. Following the same partition proposed by (Hilliard et al. 2018), we use 100/50/50 classes for training, validation and evaluation respectively. As is commonly implemented, all images are cropped and resized with the provided bounding boxes.

meta-iNat (Wertheimer and Hariharan 2019) is a fine-grained benchmark of animal species in the wild. We follow the class split proposed by where 908 classes of between 50 and 1000 images are used for training and the rest 227 are assigned for evaluation.

***tiered* meta-iNat** (Wertheimer and Hariharan 2019) is a more difficult version of meta-iNat where a large domain gap is introduced between train and test classes. We follow the same class split provided by FRN (Wertheimer, Tang, and Hariharan 2021) where 781/354 classes are used for training and evaluation respectively.

Experimental Settings

Backbone Networks. We conduct experiments on both Conv-4 and ResNet-12 backbones. Like in DN4, the Conv-4 generates a feature map of size $19 \times 19 \times 64$ (*i.e.*, 361 deep descriptors of 64 dimensions) for 84×84 image while ResNet-12 gives 25 deep descriptors of 512 dimensions.

Training and Evaluation. We meta-train Conv-4 from scratch for 30 epochs by Adam optimizer with learning rate 1×10^{-3} and decay 0.1 every 10 epochs. With regard to ResNet-12, we first pre-trained it like in the previous literature and then meta-train it by momentum SGD for 40 epochs. The learning rate in meta-training is set 5×10^{-4} for ResNet-12 and decay 0.5 every 10 epochs. We randomly sample 10,000 episodes from the test set during evaluations and compare the top-1 mean accuracy with other methods.

Re-implementations. To fully compare with recent state-of-the-art methods as well as other classic ones on *mini-/tiered*ImageNet, we re-implement ProtoNet, RelationNet, MatchingNet and baseline DN4 in a unified framework. We follow the most recent methods that take the original images as inputs instead of using the pre-cropped and resized ones in the early work. By doing so, data augmentations like random crop could be applied to achieve a fair comparison.

Method	Conv-4				ResNet-12			
	*mini*ImageNet		*tiered*ImageNet		*mini*ImageNet		*tiered*ImageNet	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MatchingNet† (Vinyals et al. 2016)	53.95	69.88	56.19	74.04	63.08	75.99	68.50	80.60
ProtoNet† (Snell, Swersky, and Zemel 2017)	52.32	69.74	53.19	72.28	62.67	77.88	68.48	83.46
RelationNet† (Sung et al. 2018)	52.12	66.90	54.33	69.95	60.97	75.12	64.71	78.41
MetaOptNet (Lee et al. 2019)	52.87	68.76	54.71	71.76	62.64	78.63	65.99	81.56
DC (Lifchitz et al. 2019)	49.84	69.64	-	-	62.53	79.77	-	-
TAPNet (Yoon, Seo, and Moon 2019)	-	-	-	-	61.65	76.36	63.08	80.26
DN4† (Li et al. 2019)	54.66	72.92	56.86	72.16	65.35	81.10	69.60	83.41
DSN$^{\nabla\bigstar}$ (Simon et al. 2020)	51.78	68.99	53.22	71.06	62.64	78.83	67.39	82.85
DeepEMD$^{\nabla\diamondsuit}$ (Zhang et al. 2020)	52.15	65.52	50.89	66.12	65.91	82.41	71.16	86.03
Negative Margin$^{\diamondsuit}$ (Liu et al. 2020)	52.84	70.41	-	-	63.85	81.57	-	-
Meta-Baseline (Chen et al. 2020)	-	-	-	-	63.17	79.26	68.62	83.29
Centroid$^{\diamondsuit}$ (Afrasiyabi, Lalonde, and Gagn'e 2020)	53.14	71.45	-	-	59.88	80.35	69.29	85.97
FEAT (Ye et al. 2020)	55.15	71.61	-	-	66.78	82.05	70.80	84.79
E^{3}BM$^{\flat\diamondsuit}$ (Liu, Schiele, and Sun 2020)	53.20	65.10	52.10	70.20	64.09	80.29	71.34	85.82
RFS-Simple (Tian et al. 2020)	55.25	71.56	56.18	72.99	62.02	79.64	69.74	84.41
RFS-Distill$^{\flat}$ (Tian et al. 2020)	**55.88**	71.65	56.76	73.21	64.82	82.14	71.52	**86.03**
FRN$^{\nabla\bigstar}$ (Wertheimer, Tang, and Hariharan 2021)	54.87	71.56	55.54	**74.68**	66.45	82.83	**72.06**	**86.89**
MN4 (ours)	55.57	**73.64**	**57.01**	73.74	**66.53**	83.39	71.95	85.66
DMN4 (ours)	**55.77**	**74.22**	**56.99**	**74.13**	**66.58**	**83.52**	**72.10**	85.72

Table 1: Few-shot classification accuracy (%) on *mini*ImageNet and *tiered*ImageNet dataset with Conv-4/ResNet-12 backbones. We show top two performances in bold font regardless of their different settings (†: our reimplementation under the same setting. ∇: the reimplemented results with their provided codes on Conv-4. $\diamondsuit$: use SGD fine-tuning during evaluation. $\flat$: knowledge distillation or model ensemble. $\bigstar$: larger shot training). The confidence intervals for our models are all below 0.25.

Few-shot Classification Results

Comparisons with the state-of-the-arts. Table 1 shows that (D)MN4 achieve new state-of-the-art with simple Conv-4 backbone and have competitive performances when using deeper ResNet-12. (D)MN4 leverages pre-training (in ResNet-12) but no other extra techniques or tricks like inference-time gradient fine-tuning, model ensembling and knowledge distillation.

Comparisons with global feature based methods. Descriptors based methods (*e.g.*, DN4, DC and DeepEMD) generally outperform classic metric-based methods that rely on the image-level feature vector (*e.g.*, MatchingNet, ProtoNet and RelationNet) by a large margin, which validates the effectiveness of using deep descriptors.

Comparisons with descriptor based methods. Among those methods, DN4 performs NBNN to represent image-level distance; DC averages predictions from each local descriptor; DeepEMD uses optimal matching to connect query and support descriptors. They all use the entire descriptor set while ignoring that some of them are not such discriminative. Our (D)MN4 outperform other model variants on almost all tasks as we think it meaningless to consider a descriptor if it is not task-relevant enough.

Comparisons with subspace methods. Table 1 shows that class-specific subspace learning (DSN) outperforms task-specific learning (TAPNet). A possible explanation is that, compared to limited class-specific subspaces, there are far more possible variants of task-specific subspaces from different class combinations, where projection matrices could easily collapse under low-data regimes. In contrast, our (D)MN4 outperforms previous methods by (1) using MNN relations to find subspaces where local characteristics are retained in their original forms comparing to matrices decomposition used in DSN; (2) using a set of deep descriptors instead of a single vector representation to avoid model collapsing.

Comparisons on fine-grained datasets. The fine-grained few-shot classification results are shown in Table 2. It can be observed that our proposed (D)MN4 are superior across the board. Interestingly, our methods achieve overwhelming performances on CUB with the simple Conv-4 backbone. The reason is that cropped and resized images in CUB have few background clutters where MNN relations can be easily established among local characteristics, *e.g.*, *eyes* and *beak*.

Ablation Study

Different Network Generated Descriptors. It can be observed in Table 1 that MN4 has greater improvement over DN4 with Conv-4 backbone compared to ResNet-12 and notice that descriptors from ResNet-12 ($r = 25, d = 640$) are much scarcer but more informative than those from Conv-4 ($r = 361, d = 64$). We speculate that different quantities and informative quality of descriptors benefit differently from the mutual nearest neighbor selectivity.

Method	CUB		meta-iNat		*tiered* meta-iNat	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet$^{\heartsuit}$ (Snell, Swersky, and Zemel 2017)	63.73	81.50	55.34	76.43	34.34	57.13
Covar. pool$^{\heartsuit}$ (Wertheimer and Hariharan 2019)	-	-	57.15	77.20	36.06	57.48
DSN$^{\heartsuit}$ (Simon et al. 2020)	66.01	85.41	58.08	77.38	36.82	60.11
CTX$^{\heartsuit}$ (Doersch, Gupta, and Zisserman 2020)	69.64	87.31	60.03	78.80	36.83	60.84
DN4† (Li et al. 2019)	73.42	90.38	62.32	79.76	43.82	64.17
FRN (Wertheimer, Tang, and Hariharan 2021)	73.48	88.43	62.42	**80.45**	43.91	63.36
MN4 (ours)	**78.10**	**92.14**	**62.87**	80.22	**43.96**	**66.93**
DMN4 (ours)	**78.36**	**92.16**	**63.00**	**80.58**	**44.10**	**67.18**

Table 2: Comparisons of 5-way few-shot classification (%) results on fine-grained datasets using Conv-4 backbone. $\heartsuit$ indicates results reported by FRN (Wertheimer, Tang, and Hariharan 2021). The confidence intervals are all below 0.25.

	(a) different network depth			(b) fix informative quality d			(c) fix descriptor quality r		
	$r = 25$ $d = 640$	$r = 100$ $d = 320$	$r = 400$ $d = 160$	$r = 25$ $d = 640$	$r = 100$ $d = 640$	$r = 400$ $d = 640$	$r = 100$ $d = 64$	$r = 100$ $d = 160$	$r = 100$ $d = 320$
DN4	65.35	61.73	57.00	65.35	62.60	63.00	59.16	61.15	61.73
MN4	66.53	62.80	58.02	66.53	63.92	64.13	59.22	61.89	62.80
DMN4	66.58	62.94	58.73	66.58	64.32	64.77	59.25	62.23	62.94

Table 3: Ablations (5-way 1-shot *mini*ImageNet tasks) on different embedding backbones derived from ResNet-12 by: (a) remove entire residual blocks to get larger number r of deep descriptors but less dimensions d; (b) remove max pooling layers within some residual blocks but fix feature dimensions d; (c) increase feature dimensions d but fix the descriptor's quantity r.

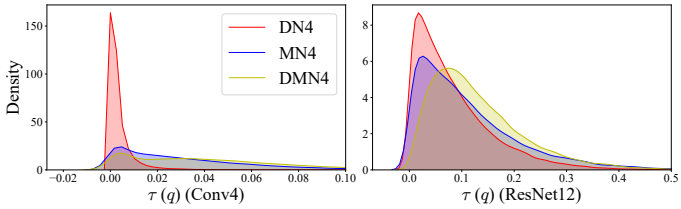


Figure 4: Kernel Density Estimation (KDE) of $\tau(q)$ from sampled query descriptors in three methods.

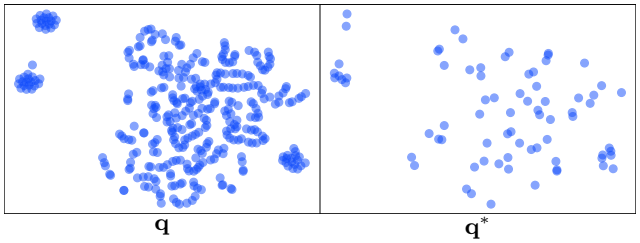


Figure 5: The t-SNE visualization of MN4 selected descriptors on an example of *mini*ImageNet with Conv-4 backbone.

To validate and further investigate the benefit of proposed methods for different kinds of network generated descriptors, we conduct ablations on different embedding backbones derived from ResNet-12. Table 3(a) firstly shows descriptors from various network depths where deeper embedding networks have less improvement with MNN. It supports the intuition that descriptors from a deep backbone own a large receptive field and contain compact image information where ignoring part of them could be helpless. Table 3(bc) shows the impact of descriptor quantity r and informative quality d by controlling variables. It can be concluded that MNN have a larger benefit when there are more deep descriptors and more information contained among them. Combining Table 3(a) with Table 3(b)(c), we can also conclude that the influence of quantity r is larger than the informative quality d of deep descriptors.

Quality of Selected Descriptors in (D)MN4. We have claimed that query descriptor $q \in \mathbf{q}^*$ that mutual nearest neighbor to some support descriptors contains class-specific information. To validate, we visualize the $\mathbf{q}$ and selected $\mathbf{q}^*$ with t-SNE in Figure 5. It can be observed that the visualization of $\mathbf{q}^*$ is departed while $\mathbf{q}$ is a mess.

We also claim (D)MN4 being able to find discriminative query descriptors in this work. To investigate the definition of such discriminability, we further recast NBNN to the log-odds updates of each class and find the relative similarity $\tau(q)$ in Eqn.(12) can be a good measure. To measure the discriminative quality of selected descriptors, we conduct the experiment by randomly sampling 50K query descriptors from DN4, MN4 and DMN4 respectively on *mini*ImageNet and visualize their kernel density estimations (KDE) of $\tau(q)$ in Figure 4. It can be found that most descriptors in DN4 contribute little in classification ($\tau(q) \approx 0$) which verifies our claim that not all descriptors are task-relevant. We also find the KDE of Conv-4 backbone is much steeper than that of ResNet-12 indicating deep compact descriptors in ResNet-12 are generally informative and useful. Overall, $\tau(q)$ are much larger in DMN4 revealing that more discriminative descriptors are selected and sampled.

	DN4 with ODM				MN4 (k%)
	30%	25%	20%	15%	
CUB	77.02	77.31	76.19	77.23	**78.10** (20.9%)
*tiered*ImageNet	56.66	56.58	56.23	55.97	**57.01** (20.0%)
*mini*ImageNet	54.40	53.76	53.99	53.73	**55.57** (25.0%)

Table 4: Classification accuracy (%) of DN4 (Conv-4) with Online Discriminative Mining (ODM) on the 5-way 1-shot tasks. We report the averaged percentage (k%) of selected query descriptors in MN4 for comparisons.

Why Still Need Mutual Nearest Neighbor? We have claimed that $\tau(q)$ is a good measure of discriminative effect for the query descriptors and our goal is to find such descriptors. Thus, it is straightforward to raise a solution by selecting the top k% query descriptors of largest $\tau(q)$ like OHEM (Shrivastava, Gupta, and Girshick 2016). To compare, we conduct an experiment by choosing top $[30\%, 25\%, 20\%, 15\%]$ query descriptors for NBNN classification and the results are shown in Table 4. It can be observed that *Online Discriminative Mining* (ODM) could definitely improve the classification accuracy, however, MN4 still outperforms them in all tasks. We speculate that this improvement is due to MNN being able to preserve *rank* (*i.e.*, variety) of selected descriptors, where aggregative query descriptors would be ignored if they are nearest neighbors to the same support descriptor but not neighbored back from it. In contrast, ODM only focuses on the top discriminative query descriptors but ignores that similar descriptors (*e.g.*, adjacent background descriptors) would be all retained if they were discriminative enough. To validate, we conduct a rank accuracy experiment by replacing the absolute similarity scores with (rank) counts:

$$\hat{c} = \arg\max_{c} \sum_{q \in \mathbf{q}^*} \mathbb{1}(c = c^*) \tag{13}$$

where $\mathbb{1}$ is an indicator function that equals 1 if its argument is true and zero otherwise. c^* is the nearest supporting class of query descriptor q as defined in Eqn.(11).

It can be observed in Table 5 that the performances of DN4 drop by a large margin if we only count the number of nearest neighbored descriptors in each support class. ODM narrows down the gap by focusing on the top discriminative descriptors. Our (D)MN4 further cuts down the differences by preserving more variety of visual characteristics with mutual nearest relations. More qualitative results in supplementary materials also demonstrate that vast majority of selected query descriptors nearest neighbor to the ground truth class.

Semi-Supervised Few-Shot Learning

From the perspective of MNN that descriptors from an unlabeled image can be roughly categorized to its MNN support class (if exists), our work can be easily extended to *semi*-supervised version MN4-semi as follows: (1) We first pseudo-label each descriptor u from unlabeled images to its MNN support class c and attach it to the support descriptors $\mathbf{s}'_c = \mathbf{s}_c \cup \{u\}$ if their MNN relationship exists. (2) We merge the support descriptor pool from all supporting descriptors and their attached pseudo-labeled descriptors $\mathbf{S}' = \bigcup_{c \in C} \mathbf{s}'_c$. (3) We run standard MN4 between $\mathbf{S}'$ and $\mathbf{q}$ as before.

We follow the same experimental setup proposed by (Ren et al. 2018) and report the comparisons in Table 6, where MN4-semi shows a consistent $\sim$ 2% improvement over the baseline DN4. Also, MN4-semi has a less performance drop compared to DSN (Simon et al. 2020) when unlabeled *distractor* classes included as MNN discards outliers from these classes.

		*mini*ImageNet		CUB	
		1-shot	5-shot	1-shot	5-shot
DN4	NBNN	54.66	72.92	73.42	90.38
	Rank	$48.84^{\downarrow 5.8}$	$58.28^{\downarrow 14.6}$	$54.80^{\downarrow 18.6}$	$77.15^{\downarrow 13.2}$
DN4 (ODM)	NBNN	54.50	72.60	76.34	92.12
	Rank	$51.42^{\downarrow 3.1}$	$69.00^{\downarrow 3.6}$	$65.93^{\downarrow 10.4}$	$88.69^{\downarrow 3.5}$
MN4	NBNN	55.57	73.64	78.10	92.28
	Rank	$52.71^{\downarrow 2.9}$	$71.53^{\downarrow 2.1}$	$\mathbf{72.20}^{\downarrow \mathbf{5.9}}$	$\mathbf{89.99}^{\downarrow \mathbf{2.1}}$
DMN4	NBNN	55.77	74.22	78.36	92.11
	Rank	$\mathbf{53.97}^{\downarrow \mathbf{2.3}}$	$\mathbf{72.60}^{\downarrow \mathbf{1.6}}$	$70.58^{\downarrow 7.8}$	$88.80^{\downarrow 3.3}$

Table 5: Comparisons of NBNN accuracy and rank accuracy (%) from three models with Conv-4 backbone. For each task, the smallest gap between NBNN and rank accuracy is marked in bold.

	Model	5-way Accuracy (%)	
		1-shot	5-shot
w/o D	PN, Non-Masked (2018)	50.09	64.59
	PN, Masked (2018)	50.41	64.39
	TPN-semi (2018)	52.78	66.42
	DSN-semi (2020)	53.01	69.12
	DN4† (2019)	**51.46**	**68.75**
	MN4-semi (ours)	**53.48**	**71.06**
w/ D	PN, Non-masked (2018)	48.70	63.55
	PN, Masked (2018)	49.04	62.96
	DSN-semi (2020)	51.01	67.12
	MN4-semi (ours)	**52.73**	**70.31**

Table 6: Semi-supervised few-shot classification results using Conv-4 on *mini*ImageNet with 40% labeled data. We show the classification results (w/ D) and without *distractors* (w/o D).

Conclusions

In this paper, we argue that not all deep descriptors are useful in recent few-shot learning methods since task-irrelevant outlier could be misleading and background descriptors could even overwhelm the object's presence. We propose Discriminative Mutual Nearest Neighbor Neural Network (DMN4) to find those that are most task-relevant to each task. Experimental results demonstrate that our method outperforms the previous state-of-the-arts on both supervised and semi-supervised FSL tasks.

Acknowledgements

This work was supported in part by The National Key Research and Development Program of China (Grant Nos: 2018AAA0101400), in part by The National Nature Science Foundation of China (Grant Nos: 62036009, U1909203, 61936006, 62133013), in part by Innovation Capability Support Program of Shaanxi (Program No. 2021TD-05).

References

Afrasiyabi, A.; Lalonde, J.-F.; and Gagn'e, C. 2020. Associative Alignment for Few-shot Image Classification. In *European Conference on Computer Vision*, 18–35. Springer.

Boiman, O.; Shechtman, E.; and Irani, M. 2008. In defense of Nearest-Neighbor based image classification. In *CVPR*.

Chen, Y.; Wang, X.; Liu, Z.; Xu, H.; and Darrell, T. 2020. A new meta-baseline for few-shot learning. *arXiv preprint arXiv:2003.04390*.

Doersch, C.; Gupta, A.; and Zisserman, A. 2020. Crosstransformers: spatially-aware few-shot transfer. *arXiv preprint arXiv:2007.11498*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 1126–1135. JMLR. org.

Gowda, K.; and Krishna, G. 1979. The condensed nearest neighbor rule using the concept of mutual nearest neighborhood (Corresp.). *IEEE Transactions on Information Theory*, 25(4): 488–490.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hilliard, N.; Phillips, L.; Howland, S.; Yankov, A.; Corley, C. D.; and Hodas, N. O. 2018. Few-Shot Learning with Metric-Agnostic Conditional Embeddings. *arXiv preprint arXiv:1802.04376*.

Huang, H.; Wu, Z.; Li, W.; Huo, J.; and Gao, Y. 2021. Local descriptor-based multi-prototype network for few-shot Learning. *Pattern Recognition*, 116: 107935.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1097–1105.

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

Li, W.; Wang, L.; Xu, J.; Huo, J.; Gao, Y.; and Luo, J. 2019. Revisiting local descriptor based image-to-class measure for few-shot learning. In *CVPR*, 7260–7268.

Lifchitz, Y.; Avrithis, Y.; Picard, S.; and Bursuc, A. 2019. Dense classification and implanting for few-shot learning. In *CVPR*, 9258–9267.

Liu, B.; Cao, Y.; Lin, Y.; Li, Q.; Zhang, Z.; Long, M.; and Hu, H. 2020. Negative margin matters: Understanding margin in few-shot classification. In *European Conference on Computer Vision*, 438–455. Springer.

Liu, Y.; Lee, J.; Park, M.; Kim, S.; Yang, E.; Hwang, S. J.; and Yang, Y. 2018. Learning to Propagate Labels: Transductive Propagation Network for Few-Shot Learning. In *International Conference on Learning Representations*.

Liu, Y.; Schiele, B.; and Sun, Q. 2020. An ensemble of epoch-wise empirical bayes for few-shot learning. In *European Conference on Computer Vision*, 404–421. Springer.

McCann, S.; and Lowe, D. G. 2012. Local naive bayes nearest neighbor for image classification. In *CVPR*, 3650–3656. IEEE.

Ren, M.; Ravi, S.; Triantafillou, E.; Snell, J.; Swersky, K.; Tenenbaum, J. B.; Larochelle, H.; and Zemel, R. S. 2018. Meta-Learning for Semi-Supervised Few-Shot Classification. In *ICLR*.

Rusu, A. A.; Rao, D.; Sygnowski, J.; Vinyals, O.; Pascanu, R.; Osindero, S.; and Hadsell, R. 2018. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*.

Sachin, R.; and Hugo, L. 2017. Optimization as a model for few-shot learning. *ICLR*.

Shrivastava, A.; Gupta, A.; and Girshick, R. 2016. Training region-based object detectors with online hard example mining. In *CVPR*, 761–769.

Simon, C.; Koniusz, P.; Nock, R.; and Harandi, M. 2020. Adaptive Subspaces for Few-Shot Learning. In *CVPR*, 4136–4145.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *NIPS*, 4077–4087.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *CVPR*, 1199–1208.

Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J. B.; and Isola, P. 2020. Rethinking few-shot image classification: a good embedding is all you need? In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 266–282. Springer.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. In *NIPS*, 3630–3638.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. *California Institute of Technology*.

Wertheimer, D.; and Hariharan, B. 2019. Few-shot learning with localization in realistic settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6558–6567.

Wertheimer, D.; Tang, L.; and Hariharan, B. 2021. Few-Shot Classification With Feature Map Reconstruction Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8012–8021.

Ye, H.-J.; Hu, H.; Zhan, D.-C.; and Sha, F. 2020. Few-shot learning via embedding adaptation with set-to-set functions. In *CVPR*, 8808–8817.

Yoon, S. W.; Seo, J.; and Moon, J. 2019. TapNet: Neural Network Augmented with Task-Adaptive Projection for Few-Shot Learning. In *ICML*, 7115–7123.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2020. DeepEMD: Few-Shot Image Classification with Differentiable Earth Mover's Distance and Structured Classifiers. In *CVPR*, 12203–12213.