

Few-shot Learning with Noisy Labels

Kevin J Liang¹ Samrudhdhi B. Rangrej² Vladan Petrovic¹ Tal Hassner¹

¹Facebook AI Research ²McGill University

kevinjliang@fb.com

Abstract

Few-shot learning (FSL) methods typically assume clean support sets with accurately labeled samples when training on novel classes. This assumption can often be unrealistic: support sets, no matter how small, can still include mislabeled samples. Robustness to label noise is therefore essential for FSL methods to be practical, but this problem surprisingly remains largely unexplored. To address mislabeled samples in FSL settings, we make several technical contributions. (1) We offer simple, yet effective, feature aggregation methods, improving the prototypes used by ProtoNet, a popular FSL technique. (2) We describe a novel Transformer model for Noisy Few-Shot Learning (TranFS). TranFS leverages a transformer’s attention mechanism to weigh mislabeled versus correct samples. (3) Finally, we extensively test these methods on noisy versions of MiniImageNet and TieredImageNet. Our results show that TranFS is on-par with leading FSL methods on clean support sets, yet outperforms them, by far, in the presence of label noise.

1. Introduction

Modern few-shot learning (FSL) methods aim to learn classifiers for novel classes from only a handful of examples. These methods, however, generally assume that the few *support set* samples used for training were carefully selected to represent their class. Unfortunately, real-world settings rarely offer such guarantees. In fact, even carefully annotated and curated datasets often contain mislabeled samples [9, 34, 39, 51, 59], due to automated weakly supervised annotation, ambiguity, or even human error.

Whereas there are plenty of methods designed for learning with noise in many-shot supervised settings [1, 20, 22, 28, 37, 57], noise in few-shot settings remains largely unexplored. This dearth is surprising considering the utility of FSL methods in settings where human supervision cannot easily be provided: *e.g.* in fully automated systems which learn many novel classes [12, 23, 49, 62, 63], making human curation of the labels of every support set, unrealistic.

Fig. 1 shows the challenge of learning from few, possibly



Figure 1. **Few-shot learning with mislabeled samples.** A 5-shot, 5-way support set of MiniImageNet [54] images. Rows show support set samples of each novel class. Two samples in each row were mislabeled by symmetric label flips (Sec. 6.1). Can you spot which ones? See Appx. A for answers and more examples.

mislabeled, examples. It presents a sample 5-shot, 5-way support set from MiniImageNet [54]. Each row includes the support set training images of one of the five classes. Two of the samples in each row are mislabeled with symmetric label noise (Sec. 6.1). With so few examples, spotting mislabeled images can be difficult, even for humans with considerable prior knowledge, which FSL methods lack.

As we later demonstrate empirically, FSL methods are especially vulnerable to such label noise. When training from few samples, each sample represents a significant contribution to the final decision boundary. Thus, even a single noisy example can be destructive to the model’s accuracy. We illustrate this observation in Fig. 2, which reports the performance of ProtoNet [46], a popular FSL method, on MiniImageNet with noisy labels. ProtoNet averages the convolutional features of each class’s support set into class *prototypes*. Queries are then classified by the class of their nearest neighbor prototype. Fig. 2 shows the effect of increasing the number of mislabeled samples, compared with a model trained after mislabeled samples were removed

(*i.e.*, smaller, but cleaner, support sets). The widening gap between the two curves reflects the degradation of accuracy when mislabeled samples are not accounted for.

We address the vulnerability of FSL methods to label noise by making a number of technical innovations. We begin by exploring simple, yet effective alternatives to the design of ProtoNet [46]. Specifically, we replace the mean operator, used by ProtoNet for aggregating support set features, with more robust methods. We evaluate an unweighted option, the median, and options which weigh support set samples based on feature similarities. We show that these changes already improve robustness to label noise.

We then introduce our **Transformer** model for **Noisy Few-Shot Learning** (TraNFS). Unlike previous methods, TraNFS *learns* to aggregate support samples into class representations. The transformer architecture offers a natural means for processing variable numbers of shots and ways with permutation invariance. Robustness to label noise is achieved by leveraging a modified version of the transformer’s self-attention mechanism [53]. This modified self-attention used by TraNFS compares support set samples and downweights samples considered likely to be mislabeled.

We test our proposed methods extensively on versions of MiniImageNet [54] and TieredImageNet [44] with three methods of adding label noise. Our results show that the proposed TraNFS (and even the simpler modifications of ProtoNet) surpass popular FSL methods by wide margins in the presence of label noise, while offering comparable performance in the absence of label noise.

To summarize, we make the following contributions.

- We propose median and similarity weighting as simple yet effective substitutes to ProtoNet’s mean prototypes.
- We present TraNFS, a novel transformer model adapted to FSL with noisy labels.
- We extensively benchmark many popular FSL methods on three types of support set noise pollution: symmetric, paired, and outlier.

2. Related work

Few-shot learning. The field of FSL methods is vast; we refer to surveys for comprehensive overviews [6, 7].

Metric-based methods classify query samples based on their similarity to each class’s support examples, learning a transferable embedding space for which such comparisons can be made. Metrics such as cosine similarity [54], Euclidean distance [46], Mahalanobis distance [6], and Earth Mover’s Distance (EMD) [65] have been shown to be effective. RelationNet [47] and Satorras *et al.* [45] used convolutional and graph neural networks, respectively, to learn a similarity metric. TADAM [40], FEAT [61], and TAFE-Net [56] proposed task-specific adaptation of embeddings. CrossTransformers [16] used attention for spatially-aware similarity between local features.

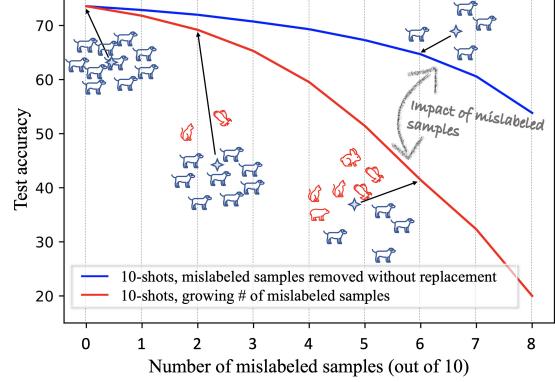


Figure 2. **Number of mislabeled samples versus few-shot learning accuracy.** Accuracy for 10-shot, 5-way classification on Mini-ImageNet [54] reported for a vanilla ProtoNet [46]. The animals represent support set embeddings, with the mean prototype (star) being pulled out of the clean class (dog) distribution with increasing mislabeled samples. Blue: Accuracy if mislabeled samples are known and ignored. Red: Accuracy when using full support sets without removing mislabeled samples. The gap between these two curves reflects the vulnerability of few-shot learning to label noise.

Optimization-based methods fine-tune model parameters on few support examples. MAML [2, 17] learned model parameter initializations that allow fast fine-tuning on few samples. REPTILE [38] simplified MAML with a first-order formulation. MetaNet [36] introduced fast and slow weights for fast parameterization and rapid generalization. Bertinetto *et al.* [8] and MetaOptNet [26] presented closed form solutions and differentiable solvers for task-dependent Ridge Regression, Logistic Regression (LR), and Support Vector Machines (SVMs). Tian *et al.* [48] showed learning of generalizable feature embeddings used to train linear classifiers on novel tasks.

Noisy labels and outliers. Methods for learning noise transition matrices are common [31, 60, 66]. Estimating a noise transition matrix from a handful of potentially mislabeled samples, however, is an ill-posed problem. Other methods [20, 22, 64] leverage deep neural networks’ tendency to learn easier (and thus likely correctly labeled) samples first [4, 50] to select reliable samples to learn from, but such behavior cannot be relied upon when only a few samples are available. Deep out-of-distribution (OOD) detection is also extensively explored [10, 15, 43, 55], but these methods typically focus on identifying test-time outliers that are out-of-distribution relative to the training set. In FSL, disjoint base and novel sets mean that *all* meta-test samples are considered OOD, including correctly labeled support set samples. Finally, there are several works that take a meta-learning approach to learning noisy labels [27, 58, 67], but these methods typically assume known label spaces with abundant data (*i.e.* many-shots), rather than our few-shot setting. With only few training samples, these methods fail.

Robust FSL. With few previous works, noisy labels have largely been ignored by FSL methods. RNNP [35] combined data augmentation with repeated applications of k -means to produce refined prototypes, but such unsupervised clustering implicitly assumes that noisy data is from one of the support set classes. RapNets [33] proposed a BiLSTM-based attentive module to overcome representation or label noise. Alternatively, RW-MAML [25] learned to weigh support samples by extending MAML to bi-bi-level optimization, but it considers the less realistic setting of mixing in OOD tasks during metatraining rather than noisy few-shot meta-test. Finally, robustness of meta-learners to adversarial attacks have also been considered [19].

3. Preliminaries

FSL classification tasks are often referred to as K -shot N -way, where N is the number of classes being learned and K is the number of labeled samples per class to learn from. These KN samples $S = \{x_1^{(1)}, x_2^{(1)}, \dots, x_{K-1}^{(N)}, x_K^{(N)}\}$ are often referred to as the *support set*. After training, unlabeled queries are classified into one of these N classes. To produce an effective classifier of novel classes C^n from few samples, FSL models typically use knowledge transfer, leveraging a dataset of *base classes* C^b with ample labeled data. It is commonly assumed that the classes in C^n are unknown in advance and thus absent from C^b (*i.e.* $C^b \cap C^n = \emptyset$). Recent FSL methods often adopt a meta-learning paradigm, simulating the desired inference-time behavior by meta-training the model with many episodes of K -shot, N -way tasks, optimizing for accuracy on Q query samples from each of the episode’s N classes.

ProtoNets. One such relevant FSL method is Prototypical Networks (ProtoNets) [46]. ProtoNets use a convolutional feature extractor \mathcal{F} to convert each sample in the support set to an embedding $h_i^{(c)} = \mathcal{F}(x_i^{(c)}) \in \mathbb{R}^D$. These embeddings are then aggregated into N class prototypes $p^{(c)}$ using a simple mean of the embeddings for each class c :

$$p^{(c)} = \frac{1}{K} \sum_i \mathcal{F}(x_i^{(c)}). \quad (1)$$

A query sample, x_q , is then classified based on the nearest prototype in embedding space:

$$y = \underset{c}{\operatorname{argmin}} d(\mathcal{F}(x_q), p^{(c)}). \quad (2)$$

Despite its simplicity, ProtoNets remain a strong baseline, and its easy implementation makes it compelling for real-world use cases at scale. Using mean to aggregate embeddings, however, implies sensitivity to mislabeled samples, especially when only few samples are provided. Indeed, as we show in Fig. 2, incorrectly labeled samples can easily degrade accuracy of the resulting classifiers. This is a symptom of the prototypes being *pulled away* from the class’s true (unknown) mean by mislabeled samples.

4. Static alternatives to the mean

Using mean as proposed by ProtoNet [46] to aggregate features is not the only way to combine embeddings into prototypes: other aggregation methods may be better suited when mislabeled samples are expected. We begin by exploring simple alternatives to the mean, intended to make prototypes more robust to mislabeled samples while maintaining accuracy if all labels are correct.

4.1. Spatial median prototypes

The median is a natural alternative to the mean in noisy settings. While order statistics like the median are well defined for scalars, this is not the case for vectors. For scalars, there is a connection between various distribution statistics (*e.g.* mean, median, mode) and minimization of the appropriate loss functions [5]. For example, empirical mean minimizes total squared error between the mean and values in the set. Similarly, empirical median minimizes total absolute error between the median and the set, so finding a median is equivalent to minimizing the total absolute error.

This minimization generalizes well to higher dimensional spaces. We thus define a cost function to be the sum of distances to embedding vectors $h_i^{(c)}$, $i \in \{1, 2, \dots, K\}$, in the set for each class, c , and find the median vector $p^{(c)}$ minimizing this cost. For brevity, we drop the class index c in derivations that follow. To make the loss differentiable at all points, we work with a smooth version of the loss, usually referred to as the *pseudo-Huber* loss:

$$\mathcal{L}(p) = \sum_{i=1}^K \left(\sqrt{\|p - h_i\|_2^2 + \epsilon^2} - \epsilon \right), \quad (3)$$

where K is the number of vectors in the set, $\|\cdot\|_2$ is an L^2 vector norm, and ϵ is a small constant.

No closed-form solution for this minimization problem exists, so we use Newton’s method for an iterative solution:

$$p(t+1) = p(t) - \mathcal{H}^{-1}(p(t)) \cdot \nabla \mathcal{L}(p(t)). \quad (4)$$

We find the gradient, $\nabla \mathcal{L}(p)$, and the Hessian matrix, $\mathcal{H}(p)$, using matrix calculus with numerator layout as,

$$\nabla \mathcal{L}(p) = \sum_{i=1}^K \frac{p - h_i}{\sqrt{\|p - h_i\|_2^2 + \epsilon^2}}, \quad (5)$$

$$\mathcal{H}(p) = \left(\sum_{i=1}^K \frac{1}{\sqrt{\|p - h_i\|_2^2 + \epsilon^2}} \right) I_{D \times D} - UU^T, \quad (6)$$

where D is the dimension of the vector space, $I_{D \times D}$ is the identity matrix, and $U = [u_1, u_2, \dots, u_K]$ is a $D \times K$ matrix formed by stacking vectors $u_i = \frac{p - h_i}{(\|p - h_i\|_2^2 + \epsilon^2)^{\frac{3}{4}}}$.

an approximation, we can neglect the second, non-diagonal term in the Hessian, in which case the iteration becomes:

$$p(t+1) = p(t) - \frac{\sum_{i=1}^K \frac{p(t)-h_i}{\sqrt{\|p(t)-h_i\|_2^2 + \epsilon^2}}}{\sum_{i=1}^K \frac{1}{\sqrt{\|p(t)-h_i\|_2^2 + \epsilon^2}}}. \quad (7)$$

Note that the choice of pseudo-Huber loss with a small constant ϵ avoids division by zero even when the median estimate falls exactly at one of the vectors in the support set. See Appx. B for additional comments on complexity.

4.2. Similarity weighted prototypes

ProtoNet-style mean aggregation uniformly weighs all shots of a class’s support set. A clear extension to this scheme is a non-uniform weighting which suppresses outliers and amplifies clean samples. Of course, if we knew which samples were mislabeled, we could remove them from the support set, but this information is typically unavailable. Instead, we can try to identify mislabeled samples based on how the support set is arranged in feature space.

Specifically, we assume that a well-trained feature extractor \mathcal{F} embeds correctly labeled samples close to one another [24], thus on average being closer in the induced metric space than any mislabeled samples. This intuition implies that the subset of correctly labeled samples is larger than any subset of mislabeled samples from a single, unrelated class. This assumption, however, is typical to many robust estimators, including, *e.g.*, the Random Sample Consensus (RANSAC) [18]. Building on this assumption, we offer the following similarity measures.

Squared Euclidean distance. This distance is the same one minimized by ProtoNets and thus a natural choice for measuring distances when attempting to identify mislabel samples. We compute the similarity score as:

$$a_i^{(c)} = -\frac{1}{K-1} \sum_{i \neq j} \|h_i^{(c)} - h_j^{(c)}\|_2^2. \quad (8)$$

Lower distance implies being closer to other support samples, so we negate the average distance for the final score.

Absolute distance. The L^2 norm can heavily penalize large distances in few feature dimensions: a large difference in only few dimensions may result in a large distance between the features, even if they share similar values in all other dimensions. We thus also consider L^1 :

$$a_i^{(c)} = -\frac{1}{K-1} \sum_{i \neq j} |h_i^{(c)} - h_j^{(c)}|. \quad (9)$$

As with the Euclidean distance, we use a factor of -1 so that lower distances produce higher scores.

Cosine similarity. While not a proper distance metric, the cosine angle between two features is a common measure of

feature similarity in the few-shot literature [11].

$$a_i^{(c)} = \frac{1}{K-1} \sum_{i \neq j} \frac{h_i^{(c)} \cdot h_j^{(c)}}{\|h_i^{(c)}\| \|h_j^{(c)}\|}. \quad (10)$$

As the inputs are normalized, cosine similarity is less sensitive to the magnitude of the embeddings.

Aggregating features with weighted similarity. Once we obtain the average distance of each feature to others in the same support set, using one of the scores above, we produce an aggregated prototype by weighting the support samples using these scores, normalizing the result with a softmax.

$$w_i^{(c)} = \frac{\exp(a_i^{(c)}/T)}{\sum_j \exp(a_j^{(c)}/T)}, \quad (11)$$

$$p^{(c)} = \sum_i w_i^{(c)} \mathcal{F}(x_i^{(c)}), \quad (12)$$

where T is a temperature term controlling the diffuseness of the softmax. As $T \rightarrow 0$, this method picks the support sample with minimum distance to the other samples as a class prototype, while as $T \rightarrow \infty$, this reduces to the mean (*i.e.* ProtoNets [46]). We choose soft-weighting of support samples, rather than top- k selection or a hard threshold, as the latter two require either knowing the number of noisy samples or threshold tuning, which may vary depending on the class or support sample distribution.

5. Learning a prototype aggregator

The aggregation methods discussed in Sec. 4, weighted or otherwise, are hard coded: They do not adjust to differences in support set feature distributions. We hypothesize that a learned mechanism that compares support set embeddings for similarity *and then refines them*, can potentially outperform these static methods. Crucially, in typical FSL settings, the number and order of support samples and classes are arbitrary. Thus, any learned alternative to the methods of Sec. 4 must process arbitrary numbers of shots or ways while remaining permutationally invariant to both.

5.1. A transformer model for noisy FSL

Given these requirements, we propose a Transformer model for Noisy Few-Shot Learning (TraNFS) (Fig. 3). Transformers are designed to process sequences of arbitrary length while offering permutation invariance. Importantly, we note that a transformer’s self-attention mechanism [53] can be leveraged to compute similarities between support set samples and naturally weigh them when aggregating them into prototypes. To this end, we concatenate the convolutional features of a support set’s samples to form an input sequence $\mathbf{h} = [h_1^{(1)}, h_2^{(1)}, \dots, h_{K-1}^{(N)}, h_K^{(N)}]$ to the transformer, \mathcal{T} . We then make the following adaptations to the transformer to enable it to process a typical FSL support set.

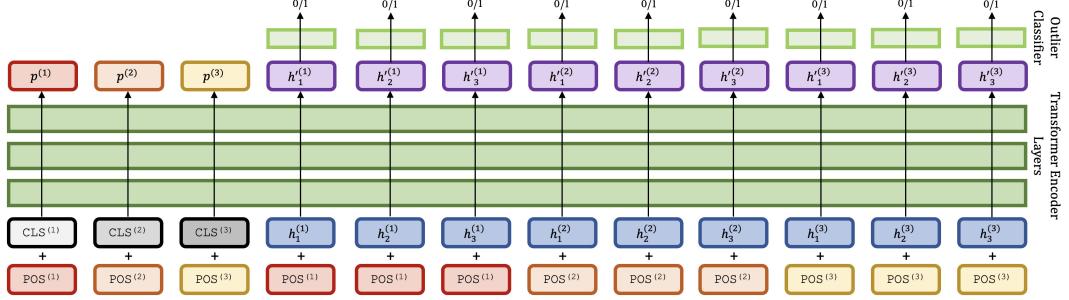


Figure 3. **Visualization of our proposed TraNFS architecture**, for a 3-shot 3-way support set example input / output sequence.

Class token. Partly inspired by BERT [14], we use a set of classification tokens $\text{CLS}^{(c)}, c \in \{1, \dots, N\}$ to denote the positions representing the prototype for each class and concatenate $[\text{CLS}^{(1)}, \dots, \text{CLS}^{(N)}]$ to the support set embedding sequence \mathbf{h} . By taking the output at the position of $\text{CLS}^{(c)}$ to be the prototype $p^{(c)}$ for class c , we motivate the transformer to learn to aggregate the information in all support set samples into this position. There are multiple choices for instantiating $\text{CLS}^{(c)}$, including as a random constant, the mean of the support set embeddings for class c (*i.e.* a mean prototype), or a learnable embedding. We report comparisons of these variations in the supplemental material.

Positional encoding. Shot and class order are both typically arbitrary in FSL and should therefore not be encoded. Still, we require some means of informing the transformer the class identity of each support sample. Vaswani *et al.* [53] utilized a sinusoidal positional encoding added to the input sequence to indicate word order. We repurpose this mechanism and use it to encode the class c associated with each position in the input sequence. Specifically, we create $N D$ -dimensional embeddings corresponding to special tokens $\text{POS}^{(c)}, c \in \{1, \dots, N\}$ and add each $\text{POS}^{(c)}$ to all support sample embeddings $h_i^{(c)}$ and the class tokens $\text{CLS}^{(c)}$, as seen in Fig. 3. With the positional embeddings added to the input sequence, the transformer can learn to attend to the positional encoding to associate support set embeddings and the prototype position of each class together.

5.2. Optimization

We meta-train TraNFS to minimize the standard ProtoNet loss. Logits are computed as the negative distance d between prototypes predicted by the model at the CLS token positions and the embedded query sample $\mathcal{F}(x_q)$:

$$\mathcal{L}_{\text{xent}} = - \sum_{c=1}^N y_q \cdot \log \left(\frac{\exp(-d(p^{(c)}, \mathcal{F}(x_q)))}{\sum_{c'} \exp(-d(p^{(c')}, \mathcal{F}(x_q)))} \right) \quad (13)$$

where \cdot is the dot product and y_q is the one-hot ground truth label of query x_q .

When meta-training TraNFS, we found it essential to expose the model to support sets with noisy samples (Sec. 6.4). We do this by artificially introducing label noise

to the support set, using label $o_i^{(c)} \in \{0, 1\}$ to track the positions of the noisy samples. This step ensures that the transformer learns a noise rejection mechanism. Without noisy samples, the transformer is not motivated to learn anything beyond recreating ProtoNet by averaging support samples.

Clean prototype loss. Besides optimizing the position of predicted prototypes relative to the meta-training query samples, we also encourage the predicted prototype for each class to be close to a *clean* prototype, $\hat{p}^{(c)}$, aggregated from correctly labeled samples in the support set:

$$\hat{p}^{(c)} = \frac{1}{K - \sum_i o_i^{(c)}} \sum_i \mathbf{1}[o_i^{(c)} = 0] \mathcal{F}(x_i^{(c)}), \quad (14)$$

$$\mathcal{L}_{\text{clean}} = \frac{1}{N} \sum_c \|p^{(c)} - \hat{p}^{(c)}\|_2^2. \quad (15)$$

We choose mean squared error here, but other alternatives such as negative cosine similarity are also viable.

Binary outlier classification loss. The ProtoNet and clean prototype losses described above both implicitly encourage identification of noisy samples. We found it helpful to also explicitly train the model to classify support set samples as either mislabeled or not.

We instantiate the binary classifier as a fully connected layer \mathcal{B} applied to the transformer’s output at positions corresponding to the support set samples. We share weights for \mathcal{B} across all such positions, with loss term:

$$\begin{aligned} \mathcal{L}_{\text{bin}} = & - \frac{1}{KN} \sum_{i,c} o_i^{(c)} \log \sigma(\mathcal{B}(h_i^{(c)})), \\ & + (1 - o_i^{(c)}) \log (1 - \sigma(\mathcal{B}(h_i^{(c)}))), \end{aligned} \quad (16)$$

where σ is the sigmoid function and $h_i^{(c)}$ is the transformer output corresponding to $h_i^{(c)}$.

Our final optimization objective combines the three losses described above:

$$\mathcal{L} = \mathcal{L}_{\text{xent}} + \lambda_c \mathcal{L}_{\text{clean}} + \lambda_b \mathcal{L}_{\text{bin}}, \quad (17)$$

where λ_c and λ_b are weighting terms for the clean prototype and binary outlier classification losses, respectively.

Table 1. **Few-shot with symmetric label swap noise.** 5-way 5-shot Acc. \pm 95% CI on MiniImageNet [54], TieredImageNet [44]. Our TraNSF is comparable to existing methods at 0% noise, with a growing gap in its favor as noise levels increase. Best viewed in color.

Model \ Noise Proportion		0%		20%		40%		60%	
Baselines	Oracle	68.18 \pm 0.16	71.42 \pm 0.18	66.08 \pm 0.17	69.19 \pm 0.19	62.60 \pm 0.17	66.14 \pm 0.20	56.89 \pm 0.18	60.39 \pm 0.21
	Nearest $k = 1$	55.91 \pm 0.17	58.81 \pm 0.20	47.27 \pm 0.18	49.48 \pm 0.19	38.68 \pm 0.18	40.25 \pm 0.19	29.20 \pm 0.16	29.84 \pm 0.17
	Nearest $k = 3$	55.29 \pm 0.18	58.44 \pm 0.20	48.43 \pm 0.17	51.11 \pm 0.19	39.14 \pm 0.17	41.09 \pm 0.18	29.66 \pm 0.15	30.69 \pm 0.15
	Nearest $k = 5$	56.15 \pm 0.18	59.22 \pm 0.20	50.92 \pm 0.17	53.75 \pm 0.19	42.12 \pm 0.17	44.14 \pm 0.19	32.62 \pm 0.16	33.99 \pm 0.17
	Linear Classifier	66.65 \pm 0.16	69.89 \pm 0.18	58.41 \pm 0.17	61.96 \pm 0.19	47.23 \pm 0.17	50.08 \pm 0.19	34.04 \pm 0.16	35.75 \pm 0.17
	Matching Networks [54]	62.16 \pm 0.17	64.92 \pm 0.19	56.21 \pm 0.18	59.20 \pm 0.20	46.18 \pm 0.18	49.12 \pm 0.20	34.66 \pm 0.18	36.80 \pm 0.19
	MAML [17]	63.25 \pm 0.18	63.96 \pm 0.19	53.28 \pm 0.18	54.62 \pm 0.19	42.58 \pm 0.18	43.71 \pm 0.19	31.01 \pm 0.17	31.74 \pm 0.17
	Vanilla ProtoNet [46]	68.27 \pm 0.16	71.36 \pm 0.18	62.43 \pm 0.17	66.15 \pm 0.19	51.41 \pm 0.19	55.05 \pm 0.21	38.33 \pm 0.19	40.61 \pm 0.21
	Baseline++ [11]	67.91 \pm 0.16	71.24 \pm 0.18	61.87 \pm 0.17	65.58 \pm 0.19	51.87 \pm 0.18	55.00 \pm 0.20	38.36 \pm 0.19	40.19 \pm 0.20
Ours	RNNP [35]	68.38 \pm 0.16	71.36 \pm 0.18	62.43 \pm 0.17	65.95 \pm 0.19	51.62 \pm 0.19	54.86 \pm 0.21	38.45 \pm 0.19	40.63 \pm 0.21
	Median	68.45 \pm 0.16	71.28 \pm 0.18	63.19 \pm 0.17	66.65 \pm 0.20	51.86 \pm 0.19	55.09 \pm 0.21	39.32 \pm 0.19	41.94 \pm 0.21
	Absolute	68.24 \pm 0.16	71.27 \pm 0.18	63.46 \pm 0.17	66.87 \pm 0.20	52.06 \pm 0.20	55.26 \pm 0.22	39.78 \pm 0.20	42.54 \pm 0.22
	Euclidean	68.32 \pm 0.16	71.48 \pm 0.18	63.02 \pm 0.17	66.69 \pm 0.19	52.09 \pm 0.19	55.62 \pm 0.21	39.33 \pm 0.20	41.75 \pm 0.21
	Cosine	68.20 \pm 0.16	70.59 \pm 0.18	63.46 \pm 0.17	66.62 \pm 0.20	52.42 \pm 0.20	55.78 \pm 0.22	39.90 \pm 0.20	42.56 \pm 0.22
TraNFS-2		68.29 \pm 0.17	70.92 \pm 0.19	64.74 \pm 0.18	67.33 \pm 0.21	56.14 \pm 0.21	58.76 \pm 0.23	42.24 \pm 0.23	44.17 \pm 0.25
TraNFS-3		68.53 \pm 0.17	71.17 \pm 0.19	65.08 \pm 0.18	67.67 \pm 0.20	56.65 \pm 0.21	58.88 \pm 0.23	42.60 \pm 0.24	44.21 \pm 0.25

6. Experiments

6.1. Experimental setup

Datasets. We experiment on two common FSL datasets: MiniImageNet [54] and TieredImageNet [44]. Both include 84×84 pixel images. MiniImageNet contains 64, 16, and 20 classes for train, validation, and test, with 60K images in total. TieredImageNet consists of 351, 97, and 160 classes for train, validation, and test, with ~ 0.78 M images in total.

Label noise types. We explore three forms of label noise:

- *Symmetric label swap* noise [52] draws mislabeled samples, uniformly at random, from the other $N - 1$ classes of the episode, with the restriction that a noisy class does not tie or outnumber the original clean class.
- *Paired label swap* noise [20] is more challenging: we always draw mislabeled samples from the same class by assigning each class with a *noisy class* counterpart, simulating real-world tendencies to confuse certain classes with others during labeling. We randomly generate these assignments in each episode as a derangement, to prevent models from learning these pairings across episodes.
- *Outlier* noise is sampled from classes *outside* the N -way episode. We use 600 images from each of the 350 ImageNet classes unincluded from MiniImageNet and TieredImageNet. We split these classes in half for meta-training and meta-test, to ensure that meta-test episode outliers represent previously unseen classes.

The amount of noise in a support set is specified as the percent of the total number of shots. We only consider settings where the clean class can reasonably be identified. Thus, for example, we only consider paired label swap noise under 50%, as at 50% noise and above, the clean class is ambiguous or a minority. We also exclude paired label swap settings that are identical to the corresponding symmetric label swap setting (*e.g.* 20% for 5-shot 5-way).

Table 2. **Few-shot with paired label swap noise.** 5-way 5-shot Acc. \pm 95% CI on MiniImageNet [54], TieredImageNet [44].

Model \ Noise Proportion		40%	
Baselines	Oracle	62.60 \pm 0.17	66.14 \pm 0.20
	Nearest $k = 1$	37.97 \pm 0.17	39.40 \pm 0.18
	Nearest $k = 3$	37.84 \pm 0.16	39.70 \pm 0.18
	Nearest $k = 5$	40.39 \pm 0.17	42.17 \pm 0.18
	Linear Classifier	44.49 \pm 0.17	46.70 \pm 0.18
	Matching Networks [54]	43.53 \pm 0.17	46.13 \pm 0.19
	MAML [17]	40.67 \pm 0.18	41.66 \pm 0.18
	Vanilla ProtoNet [46]	47.77 \pm 0.19	50.85 \pm 0.21
	Baseline++ [11]	47.82 \pm 0.18	50.69 \pm 0.20
Ours	RNNP [35]	47.88 \pm 0.19	50.91 \pm 0.20
	Median	48.81 \pm 0.19	51.91 \pm 0.21
	Absolute	49.38 \pm 0.20	52.40 \pm 0.22
	Euclidean	48.67 \pm 0.19	51.90 \pm 0.21
	Cosine	49.40 \pm 0.19	52.72 \pm 0.22
TranFS-2		50.63 \pm 0.22	54.82 \pm 0.24
TranFS-3		53.96 \pm 0.23	55.12 \pm 0.24

Model. Our models are implemented in PyTorch [41], using learn2learn [3] as a starting point. We set $T = 25$ for similarity weighted prototypes with squared euclidean and absolute distances and $T = 0.2$ for cosine similarity. For TranFS, we instantiate the transformer \mathcal{T} using 2 or 3-layers with eight heads, learnable positional embeddings $\text{POS}^{(c)}$, and random constant class tokens $\text{CLS}^{(c)}$. We apply an orthogonally initialized pair of down-projection and up-projection weight matrices before and after the transformer, reducing the transformer’s dimensionality to 128. We found these orthogonal projections stabilize training [42], while also greatly reducing the number of transformer parameters. Finally, we set the hyperparameters of Eq. (17) as $\lambda_b = 0.5$ and $\lambda_c = 5$. See Appx. E for hyperparameter sweeps. Note that while the transformer must be meta-trained and used to generate robust prototypes during meta-test, it is not used during inference on individual query samples. Hence, the number of parameters and computation cost during inference is similar to methods like ProtoNet [46].

Training and testing. To isolate the effect of the method

Table 3. **Few-shot with outlier noise.** 5-way 5-shot Acc. \pm 95% CI on MiniImageNet [54], TieredImageNet [44]. Our TraNSF is comparable to existing methods at 0% or low noise, with a growing gap in its favor as noise levels increase. Best viewed in color.

Model \ Noise Proportion		0%		20%		40%		60%	
Baselines	Oracle	68.18 \pm 0.16	71.42 \pm 0.18	66.08 \pm 0.17	69.19 \pm 0.19	62.60 \pm 0.17	66.14 \pm 0.20	56.89 \pm 0.18	60.39 \pm 0.21
	Nearest $k = 1$	55.87 \pm 0.18	58.89 \pm 0.20	50.90 \pm 0.18	54.57 \pm 0.20	45.28 \pm 0.18	49.45 \pm 0.20	38.75 \pm 0.18	43.20 \pm 0.19
	Nearest $k = 3$	55.28 \pm 0.18	58.38 \pm 0.20	50.53 \pm 0.17	53.98 \pm 0.20	44.40 \pm 0.17	48.06 \pm 0.19	37.03 \pm 0.16	40.11 \pm 0.18
	Nearest $k = 5$	56.34 \pm 0.17	59.25 \pm 0.19	52.32 \pm 0.17	55.30 \pm 0.19	46.49 \pm 0.17	49.34 \pm 0.19	38.44 \pm 0.16	40.56 \pm 0.17
	Linear Classifier	66.70 \pm 0.16	69.60 \pm 0.18	61.13 \pm 0.17	64.58 \pm 0.19	53.86 \pm 0.18	57.57 \pm 0.20	44.05 \pm 0.18	47.90 \pm 0.20
	Matching Networks [54]	62.05 \pm 0.17	64.99 \pm 0.19	57.69 \pm 0.18	60.74 \pm 0.20	51.32 \pm 0.19	54.28 \pm 0.21	42.39 \pm 0.19	44.93 \pm 0.20
	MAML [17]	63.21 \pm 0.18	63.90 \pm 0.19	57.35 \pm 0.19	58.14 \pm 0.19	50.00 \pm 0.19	51.11 \pm 0.20	40.90 \pm 0.17	42.01 \pm 0.20
	Vanilla ProtoNet [46]	68.18 \pm 0.16	71.42 \pm 0.18	63.92 \pm 0.17	67.58 \pm 0.19	57.07 \pm 0.18	60.97 \pm 0.20	46.99 \pm 0.20	50.29 \pm 0.21
	Baseline++ [11]	67.85 \pm 0.16	71.29 \pm 0.18	63.49 \pm 0.17	67.07 \pm 0.19	56.84 \pm 0.18	60.64 \pm 0.20	46.96 \pm 0.19	50.07 \pm 0.21
Ours	RNNP [35]	68.17 \pm 0.16	71.28 \pm 0.18	63.80 \pm 0.17	67.29 \pm 0.19	56.97 \pm 0.18	60.83 \pm 0.20	46.92 \pm 0.20	50.09 \pm 0.21
	Median	68.37 \pm 0.16	71.28 \pm 0.18	64.46 \pm 0.17	67.79 \pm 0.19	57.85 \pm 0.18	61.63 \pm 0.21	47.19 \pm 0.20	50.63 \pm 0.21
	Absolute	68.13 \pm 0.16	71.17 \pm 0.18	64.69 \pm 0.17	68.00 \pm 0.19	58.30 \pm 0.18	61.98 \pm 0.21	47.39 \pm 0.20	50.59 \pm 0.22
	Euclidean	68.51 \pm 0.16	71.28 \pm 0.18	64.57 \pm 0.17	67.89 \pm 0.19	58.01 \pm 0.18	61.61 \pm 0.20	47.25 \pm 0.20	50.49 \pm 0.21
Ours	Cosine	68.20 \pm 0.16	70.79 \pm 0.18	64.78 \pm 0.17	67.94 \pm 0.19	58.36 \pm 0.18	62.37 \pm 0.21	47.34 \pm 0.20	51.12 \pm 0.22
	TraNFS-2	67.76 \pm 0.17	70.83 \pm 0.19	64.47 \pm 0.19	67.52 \pm 0.21	58.29 \pm 0.20	61.76 \pm 0.22	47.37 \pm 0.23	51.40 \pm 0.23
	TraNFS-3	68.11 \pm 0.17	71.13 \pm 0.19	64.96 \pm 0.18	67.93 \pm 0.20	59.03 \pm 0.20	62.39 \pm 0.22	47.69 \pm 0.22	51.82 \pm 0.23

from the learned features, we use the same, frozen, 4-layer convolutional backbone [54], trained with the ProtoNet objective, for all models except MAML [17]. We chose this simple backbone to emphasize the method, not the feature extractor. The backbone is trained with AdamW [32], with weight decay of 0.01, initial learning rate 1×10^{-3} , and learning rate decay of $\times 0.7$ every 10K episodes for 100K episodes for MiniImageNet, and every 25K episodes for 250K episodes for TieredImageNet. Meta-validation is used for accuracy model selection. Our TraNFS is similarly optimized, with initial learning rate 5×10^{-4} , decayed after every 25K episodes, for 200K episodes. We use random horizontal flips, resized crops, and color jitters as data augmentations for all models. Finally, each meta-train and meta-test episode has 15 queries. We report mean accuracy and 95% confidence interval for 10K meta-test episodes. All experiments are run on a single Nvidia V100 GPU.

6.2. Noisy few-shot results

We compare all our proposed methods for noisy few-shot learning—Median, Absolute, Euclidean, Cosine, and TraNFS—with several baselines (see Appx. D for details on these baselines). We report results for 5-way and 5-shot¹ on MiniImageNet and TieredImageNet using symmetric noise (Table 1), paired noise (Table 2), and outlier noise (Table 3). We further report results for an *Oracle*: a ProtoNet [46] that knows which samples are mislabeled and ignores them by removing them from each support set, thereby representing perfect noise rejection (the blue line in Fig. 2).

Unsurprisingly, noisy labels negatively affect all methods. Our proposed median and similarity weighting alternatives to ProtoNet’s mean suffer less than the baselines, on all three noise types. This is expected, as their aggregation methods are less sensitive to outliers. Furthermore, our transformer-based TraNFS is clearly superior to its baselines. For example, consider the challenging 5-way

¹See Appx. C for 3-shot and 10-shot experiments.

Table 4. **Various amounts of injected meta-training artificial noise.** TraNFS-3 5-way 5-shot Acc. \pm 95% CI on MiniImageNet.

0%	20%	40%	60%	0%	20%	40%	60%
✓				69.10 \pm 0.16	63.56 \pm 0.18	52.85 \pm 0.19	39.19 \pm 0.21
	✓			68.67 \pm 0.17	64.85 \pm 0.18	55.76 \pm 0.21	41.73 \pm 0.23
		✓		67.37 \pm 0.17	63.97 \pm 0.19	55.65 \pm 0.21	41.63 \pm 0.24
			✓	50.40 \pm 0.19	48.26 \pm 0.19	43.11 \pm 0.21	35.44 \pm 0.23
	✓	✓		68.53 \pm 0.17	65.08 \pm 0.18	56.65 \pm 0.21	42.60 \pm 0.24
	✓	✓	✓	68.90 \pm 0.17	65.08 \pm 0.18	56.73 \pm 0.21	42.69 \pm 0.24
	✓	✓	✓	66.92 \pm 0.17	63.52 \pm 0.19	54.98 \pm 0.22	42.01 \pm 0.24
	✓	✓	✓	67.64 \pm 0.17	63.83 \pm 0.18	54.81 \pm 0.21	41.33 \pm 0.24

Table 5. **Meta-training mean/median prototype models with noise.** 5-way 5-shot Acc. \pm 95% CI on MiniImageNet.

Baseline \ Noise Proportion	0%	20%	40%	60%
Mean + Sym (0%, 20%, 40%)	67.89 \pm 0.16	62.44 \pm 0.18	51.66 \pm 0.19	38.53 \pm 0.20
Mean + Pair (0%, 20%, 40%)	-	-	48.05 \pm 0.19	-
Mean + Out (0%, 20%, 40%)	66.88 \pm 0.16	62.69 \pm 0.17	56.00 \pm 0.18	45.90 \pm 0.19
Median + Sym (0%, 20%, 40%)	67.11 \pm 0.16	62.39 \pm 0.17	51.70 \pm 0.19	39.57 \pm 0.20
Median + Pair (0%, 20%, 40%)	-	-	48.64 \pm 0.19	-
Median + Out (0%, 20%, 40%)	67.17 \pm 0.16	63.46 \pm 0.17	57.01 \pm 0.18	46.44 \pm 0.20

5-shot setting on MiniImageNet with 40% paired label swap noise. Our TraNFS provides a 6.19% absolute improvement in accuracy over ProtoNet, representing a *significant relative drop of 41.7% in error*, compared with the Oracle. As explained in Sec. 6.3, this gain is due to the transformer’s self-attention learning to compare support set examples and suppress samples suspected of being mislabeled.

Comparing noise types, we find that, as expected, FSL methods are more vulnerable to paired label swap than symmetric noise. Outlier noise has the least impact on model performance, with three outlier samples in a 5-way 5-shot test reducing accuracy similarly to two label swap samples. We reason that this is due to the direction in which these noisy samples push the model’s decision boundary: label swapped samples pull the decision boundary closer to the features of other classes in the N -way classification task; for paired label swaps, this effort is coordinated across noisy samples, amplifying the effect. In contrast, outlier samples have lower probability of being arranged in regions of feature space that interfere with the N -way classification.

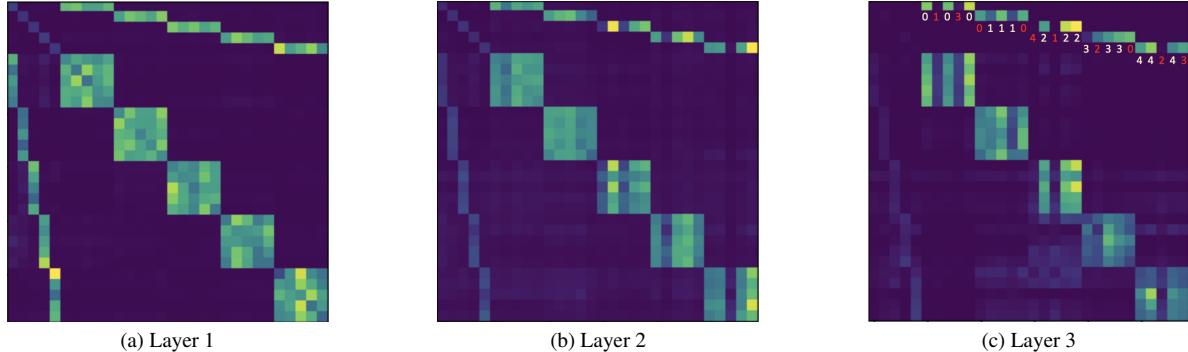


Figure 4. **Attention maps from selected TraNFS-3 self-attention heads.** 5-way 5-shot MiniImageNet, 40% symmetric noise. The first five rows of each map correspond to positions of CLS tokens, which produce the output prototypes. True class label of each support sample overlaid on the final layer’s attention maps. Evidently, later layers assign lower attention to noisy samples, effectively filtering them.

6.3. Visualizing transformer attention to noise

To understand how our proposed transformer model suppresses noisy samples, we visualize the self-attention from selected attention heads at each of its layers. The attention maps for a 5-way 5-shot test episode of MiniImageNet with 40% symmetric label swap noise are shown in Fig. 4. From Fig. 4a, attention at the CLS token positions suggests that the first layer of the transformer mainly uses positional encodings to focus on per-class examples, resulting in a representation reminiscent of ProtoNet’s mean.

Subsequent layers are visualized in Fig. 4b-4c. Evidently, the transformer is able to refine its class representations by decreasing attention to noisy samples, suppressing their influence on the aggregated representations. For example, our self-attention mechanism correctly learned to ignore the 2nd and 4th samples of the first class, which are indeed mislabeled. While this filtering ability is not perfect, we emphasize that learning class concepts from so few samples, without any prior concept of the class, is a challenging task (see Fig. 1); ImageNet contains enough intra-class variation and label ambiguity that identifying mislabeled samples can be challenging even for humans, who have the advantage of conceptual priors of the ImageNet classes.

6.4. Ablations: Meta-training noise proportion

To test the influence of adding training noise,² we meta-train a 3-layer TraNFS model on 5-shot, 5-way MiniImageNet with various amounts of symmetric label swap noise, synthetically added to meta-training. Table 4 reports these results, clearly showing a few patterns.

First, training with a single noise percentage boosts performance on that noise level during meta-test. Those models, however, do not generalize well to other noise levels. Instead, training on varying noise levels seems to offer the best results across a range of meta-test support set noise levels. This is important, as the stochastic nature of label noise means we expect real world support sets to have varying levels of noise, and it is desirable to handle multiple noise

²See Appx. E for more ablation studies on other design choices.

levels with a single model. In particular, training on support sets with {0, 20, 40}% appears to achieve the best overall performance. Finally, we observe that training on extremely noisy support sets (*e.g.*, 60%) appears counter-productive. We believe that this is due to a mixture of having a more challenging task to learn while also diluting learnable information of the clean class of each support set.

Synthetically adding noise during meta-training proved essential for TraNFS. A similar strategy could conceivably also be applied to other methods. As Table 5 shows, however, this approach was unhelpful. We believe that the absence of learnable mechanisms for rejecting noise encourages these baselines to learn strong feature extractors on the highest quality (*i.e.*, cleanest) data available. Thus, we do not add artificial noise to our baselines during meta-training.

7. Conclusion

We focus on a key vulnerability of modern FSL methods: noisy, mislabeled support sets samples. We propose several technical novelties to mitigate this vulnerability: replacing the mean aggregator used by ProtoNets with a median or similarity weighted aggregation. We then present a novel, transformer-based model designed to learn a dynamic noise rejection mechanism, leveraging the transformer’s attention mechanism. Experiments on MiniImageNet and TieredImageNet under varying types and levels of noise clearly show the effectiveness of our techniques.

Limitations. As with other FSL methods, we assume a cleanly labeled meta-train dataset with noisy labels mostly affecting meta-testing. We believe this is reasonable: collecting meta-train data and meta-training are performed offline, before model deployment, with reasonable control over both. Sometimes, however, meta-training datasets can also be noisy. While we introduce noise during meta-training, our methods assume that the query set is correctly labeled. Queries with noisy labels could cause misleading gradients. In such cases, ideas from the noisy label literature [20, 22] could offer promising direction for future work.

References

- [1] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 1988. 1
- [2] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. *arXiv preprint arXiv:1810.09502*, 2018. 2
- [3] Sébastien MR Arnold, Praateek Mahajan, Debajyoti Datta, Ian Bunner, and Konstantinos Saitas Zarkias. learn2learn: A library for meta-learning research. *arXiv preprint arXiv:2008.12284*, 2020. 6
- [4] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *Int. Conf. Mach. Learning.*, 2017. 2
- [5] Jonathan T. Barron. A general and adaptive robust loss function. In *Conf. Comput. Vis. Pattern Recog.*, 2019. 3
- [6] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved few-shot visual classification. In *Conf. Comput. Vis. Pattern Recog.*, 2020. 2
- [7] Nihar Bende, Hugo Terashima Marín, and Peyman Najafirad. Learning from few samples: A survey. *arXiv preprint arXiv:2007.15484*, 2020. 2
- [8] Luca Bertinetto, Joao F Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *Int. Conf. Learn. Represent.*, 2018. 2
- [9] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 1
- [10] Koby Bibas, Meir Feder, and Tal Hassner. Single layer predictive normalized maximum likelihood for out-of-distribution detection. In *Adv. Neural Inform. Process. Syst.*, 2021. 2
- [11] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019. 4, 6, 7, 15, 16
- [12] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. The youtube video recommendation system. In *ACM Conference on Recommender Systems*, 2010. 1
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conf. Comput. Vis. Pattern Recog.*, 2009. 12
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 5
- [15] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 2
- [16] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. In *Adv. Neural Inform. Process. Syst.*, 2020. 2
- [17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Int. Conf. Mach. Learning.*, 2017. 2, 6, 7, 12, 13, 14, 15
- [18] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981. 4
- [19] Micah Goldblum, Liam Fowl, and Tom Goldstein. Adversarially robust few-shot learning: A meta-learning approach. In *Adv. Neural Inform. Process. Syst.*, 2020. 3
- [20] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Adv. Neural Inform. Process. Syst.*, 2018. 1, 2, 6, 8
- [21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 18
- [22] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Int. Conf. Mach. Learning.*, 2018. 1, 2, 8
- [23] Dotan Kaufman, Koby Bibas, Eran Borenstein, Michael Chertok, and Tal Hassner. Balancing specialization, generalization, and compression for detection and tracking. In *Brit. Mach. Vis. Conf.*, 2019. 1
- [24] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Adv. Neural Inform. Process. Syst.*, 2020. 4
- [25] Krishnateja Killamsetty, Changbin Li, Chen Zhao, Rishabh Iyer, and Feng Chen. A reweighted meta learning framework for robust few shot learning. *arXiv preprint arXiv:2011.06782*, 2020. 3
- [26] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Conf. Comput. Vis. Pattern Recog.*, 2019. 2
- [27] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *Conf. Comput. Vis. Pattern Recog.*, 2019. 2
- [28] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Int. Conf. Comput. Vis.*, 2017. 1
- [29] Kevin J Liang, Weituo Hao, Dinghan Shen, Yufan Zhou, Weizhu Chen, Changyou Chen, and Lawrence Carin. MixKD: Towards efficient distillation of large-scale language models. In *Int. Conf. Learn. Represent.*, 2021. 18
- [30] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Ming-sheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *Eur. Conf. Comput. Vis.*, 2020. 16
- [31] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015. 2
- [32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Int. Conf. Learn. Represent.*, 2018. 7, 15

- [33] Jiang Lu, Sheng Jin, Jian Liang, and Changshui Zhang. Robust few-shot learning for user-provided data. *IEEE trans. on neural networks and learning systems*, 2020. 3
- [34] Iacopo Masi, Anh Tuan Tran, Tal Hassner, Gozde Sahin, and Gérard Medioni. Face-specific data augmentation for unconstrained face recognition. *Int. J. Comput. Vis.*, 2019. 1
- [35] Pratik Mazumder, Pravendra Singh, and Vinay P Namboodiri. Rnnlp: A robust few-shot learning approach. In *Proc. Winter Conf. on Applications of Comput. Vis.*, 2021. 3, 6, 7, 12, 13, 14, 15, 16
- [36] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *Int. Conf. Mach. Learning.*, 2017. 2
- [37] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Adv. Neural Inform. Process. Syst.*, 2013. 1
- [38] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 2
- [39] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Perseverative label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021. 1
- [40] Boris N Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Adv. Neural Inform. Process. Syst.*, 2018. 2
- [41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst.*, 2019. 6
- [42] Samrudhdhi Bharatkumar Rangrej, Kevin J Liang, Xi Yin, Guan Pang, Theofanis Karaletsos, Lior Wolf, and Tal Hassner. Revisiting linear decision boundaries for few-shot learning with transformer hypernetworks, 2021. 6
- [43] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A DePristo, Joshua V Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *arXiv preprint arXiv:1906.02845*, 2019. 2
- [44] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. In *Int. Conf. Learn. Represent.*, 2018. 2, 6, 7, 12, 16, 17
- [45] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *Int. Conf. Learn. Represent.*, 2018. 2
- [46] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Adv. Neural Inform. Process. Syst.*, 2017. 1, 2, 3, 4, 6, 7, 12, 13, 14, 15, 16
- [47] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Conf. Comput. Vis. Pattern Recog.*, 2018. 2
- [48] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Eur. Conf. Comput. Vis.*, 2020. 2
- [49] Raciel Yera Toledo, Yailé Caballero Mota, and Luis Martínez. Correcting noisy ratings in collaborative recommender systems. *Knowledge-Based Systems*, 2015. 1
- [50] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. In *Int. Conf. Learn. Represent.*, 2019. 2
- [51] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. In *Int. Conf. Mach. Learning.*, 2020. 1
- [52] Brendan Van Rooyen, Aditya Krishna Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. *arXiv preprint arXiv:1505.07634*, 2015. 6, 12
- [53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, 2017. 2, 4, 5
- [54] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Adv. Neural Inform. Process. Syst.*, 2016. 1, 2, 6, 7, 12, 13, 14, 15, 16, 17, 18
- [55] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L. Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Eur. Conf. Comput. Vis.*, 2018. 2
- [56] Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, and Joseph E Gonzalez. Tafe-net: Task-aware feature embeddings for low shot learning. In *Conf. Comput. Vis. Pattern Recog.*, 2019. 2
- [57] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *Conf. Comput. Vis. Pattern Recog.*, 2018. 1
- [58] Zhen Wang, Guosheng Hu, and Qinghua Hu. Training noise-robust deep neural networks via meta-learning. In *Conf. Comput. Vis. Pattern Recog.*, 2020. 2
- [59] Yuewei Yang, Kevin J Liang, and Lawrence Carin. Object detection as a positive-unlabeled problem. In *Brit. Mach. Vis. Conf.*, 2020. 1
- [60] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. *arXiv preprint arXiv:2006.07805*, 2020. 2
- [61] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Conf. Comput. Vis. Pattern Recog.*, 2020. 2
- [62] Mang Ye and Pong C Yuen. PurifyNet: A robust person re-identification model with noisy labels. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 1
- [63] Li Yin, Juan-Manuel Perez-Rua, and Kevin J Liang. Sylph: A hypernet-framework for incremental few-shot object detection. In *Conf. Comput. Vis. Pattern Recog.*, 2022. 1
- [64] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help gener-

- alization against label corruption? In *Int. Conf. Mach. Learning.*, 2019. 2
- [65] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover’s distance and structured classifiers. In *Conf. Comput. Vis. Pattern Recog.*, 2020. 2
- [66] Yivan Zhang, Gang Niu, and Masashi Sugiyama. Learning noise transition matrix from only noisy labels via total variation regularization. *arXiv preprint arXiv:2102.02414*, 2021. 2
- [67] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. Meta label correction for noisy label learning. In *Proc. of the AAAI Conf. on Artificial Intelligence*, 2021. 2

We include supplemental material for our work here. Appx. A shows the mislabeled samples in the 5-way 5-shot support set in Fig. 1, as well as two more example noisy support sets. We discuss computational complexity considerations for iteratively solving for the median in Sec. B. In Appx. C, we investigate noisy few-shot performance for different numbers of shots from the 5-shot setting considered in Sec. 6.2. Appx. D contains descriptions and additional implementation details of the baselines that we compare against. We perform further ablation studies beyond Sec. 6.4, investigating feature extractors, hyperparameter settings, and architectural design choices of TraNFS in Appx. E.

A. Noisy support set examples

Noisy few-shot learning is a challenging problem. Even before adding noise, there can be significant variation within a class, largely due to the manner in which the ImageNet [13] dataset (from which MiniImageNet [54] and TieredImageNet [44] are derived) was constructed. Some images in the *clean* version of ImageNet are mislabeled due to human error, but even among the correctly labeled objects, there are non-canonical views, images with multiple objects (possibly from multiple ImageNet classes), and classes that are close to synonymous. We provide several examples of noisy support sets from MiniImageNet with 40% symmetric label swap noise [52] in Fig. 5, with the clean and noisy samples framed in green and red, respectively. While humans are generally able to separate the noisy samples from the clean samples with some scrutiny, this is in large part due to prior conceptual understandings of the classes depicted. Few-shot models presented with support sets such as those in Fig. 5 are tasked with learning how to distinguish the depicted classes *without having previously seen these classes*, a much more difficult problem.

B. A Note On Median Complexity

As discussed in Sec. 4.1, median computation has to be performed iteratively since no closed form solution exists. We have chosen the 2nd- over 1st-order optimization as the former provides an optimal step size at each iteration, speeding up convergence. This choice may seem costly at first glance, but computational complexity analysis of Eq. (7) shows negligible 2nd-order method overhead. Each median update iteration takes $4DK + 2K - D$ flops for gradient computation, K flops for optimal step calculation (2nd-order method overhead), and $2D$ flops for parameter update. We emphasize that this optimization is done to calculate a median prototype (as opposed to updating the model weights); D and K are both fairly small.

Table 6. Few-shot performance with symmetric label swap noise on 5-way 3-shot MiniImageNet [54].

Model \ Noise Proportion	0%	33.3%
Oracle	62.60 ± 0.17	56.89 ± 0.18
Nearest $k = 1$	52.98 ± 0.18	39.92 ± 0.18
Nearest $k = 3$	50.59 ± 0.18	38.76 ± 0.16
Nearest $k = 5$	50.20 ± 0.17	40.05 ± 0.16
Linear Classifier	61.54 ± 0.17	46.06 ± 0.17
Matching Networks [54]	57.86 ± 0.18	44.92 ± 0.18
MAML [17]	59.79 ± 0.20	40.41 ± 0.17
Vanilla ProtoNet [46]	62.54 ± 0.18	48.78 ± 0.19
RNNP [35]	62.57 ± 0.17	48.76 ± 0.19
Median	62.60 ± 0.17	50.40 ± 0.19
Absolute $T = 10.0$	61.77 ± 0.17	50.93 ± 0.19
Absolute $T = 25.0$	62.54 ± 0.17	50.84 ± 0.19
Absolute $T = 50.0$	62.69 ± 0.17	50.06 ± 0.19
Euclidean $T = 10.0$	62.58 ± 0.17	50.83 ± 0.19
Euclidean $T = 25.0$	62.62 ± 0.18	50.06 ± 0.19
Euclidean $T = 50.0$	62.62 ± 0.17	49.51 ± 0.19
Cosine $T = 0.2$	62.75 ± 0.17	49.63 ± 0.19
Cosine $T = 0.5$	62.55 ± 0.17	49.15 ± 0.19
Cosine $T = 1.0$	62.52 ± 0.17	49.20 ± 0.19
Cosine $T = 2.0$	62.63 ± 0.17	49.05 ± 0.19
Cosine $T = 5.0$	62.54 ± 0.17	48.96 ± 0.19
TraNFS-2	64.17 ± 0.18	53.35 ± 0.21
TraNFS-3	64.28 ± 0.18	53.84 ± 0.21

Table 7. Few-shot performance with outlier noise on 5-way 3-shot MiniImageNet [54].

Model \ Noise Proportion	0%	33.3%
Oracle	62.60 ± 0.17	56.89 ± 0.18
Nearest $k = 1$	53.07 ± 0.18	44.66 ± 0.18
Nearest $k = 3$	50.40 ± 0.18	41.59 ± 0.17
Nearest $k = 5$	50.24 ± 0.17	42.26 ± 0.17
Linear Classifier	61.58 ± 0.17	51.21 ± 0.18
Matching Networks [54]	57.82 ± 0.18	48.56 ± 0.19
MAML [17]	59.76 ± 0.19	47.08 ± 0.19
Vanilla ProtoNet [46]	62.43 ± 0.17	52.78 ± 0.19
RNNP [35]	62.55 ± 0.17	52.88 ± 0.19
Median	62.53 ± 0.17	53.82 ± 0.19
Absolute $T = 10.0$	61.54 ± 0.17	53.76 ± 0.19
Absolute $T = 25.0$	62.47 ± 0.17	54.07 ± 0.19
Absolute $T = 50.0$	62.69 ± 0.17	53.73 ± 0.19
Euclidean $T = 10.0$	62.56 ± 0.18	54.10 ± 0.19
Euclidean $T = 25.0$	62.57 ± 0.17	53.72 ± 0.18
Euclidean $T = 50.0$	62.76 ± 0.17	53.55 ± 0.19
Cosine $T = 0.2$	62.58 ± 0.17	53.46 ± 0.19
Cosine $T = 0.5$	62.50 ± 0.17	53.03 ± 0.19
Cosine $T = 1.0$	62.50 ± 0.17	52.84 ± 0.19
Cosine $T = 2.0$	62.72 ± 0.17	53.16 ± 0.19
Cosine $T = 5.0$	62.63 ± 0.18	53.19 ± 0.19
TraNFS-2	63.63 ± 0.18	54.75 ± 0.20
TraNFS-3	63.61 ± 0.18	54.72 ± 0.20

C. Different number of shots

While the experiments in Sec. 6.2 are conducted with 5 shots, many of our findings on noisy FSL apply to other numbers of shots as well. We provide additional results below for MiniImageNet [54] with $K = \{3, 10\}$ shots.

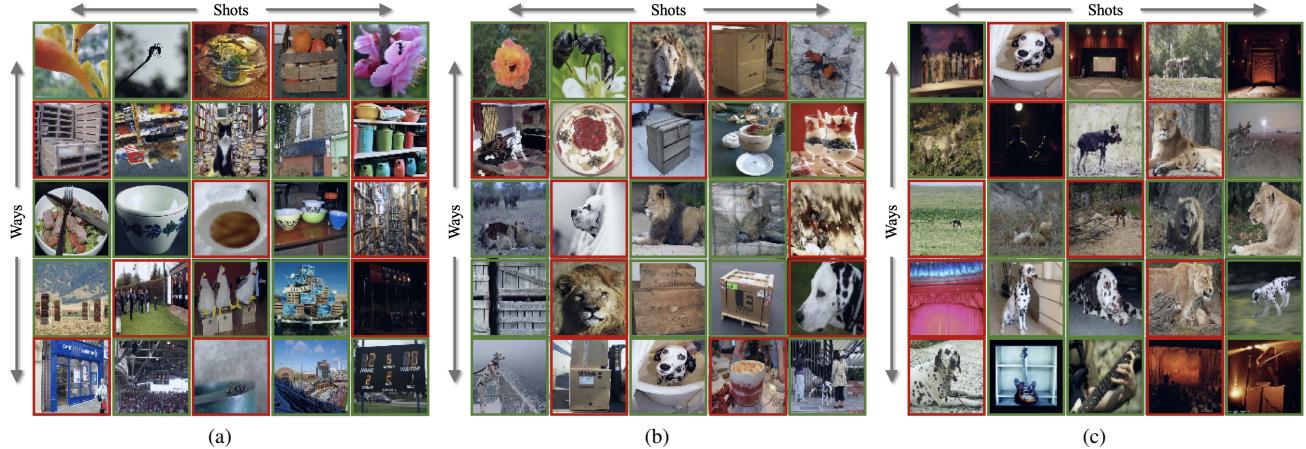


Figure 5. **Noisy support set examples.** Images with green boxes are clean samples from the original class, while red boxes are mislabeled samples due to symmetric label swaps. (a) is the support set shown in Fig. 1 of the main paper.

Table 8. Few-shot performance with symmetric label swap noise on 5-way 10-shot MiniImageNet [54].

Model \ Noise Proportion	0%	10%	20%	30%	40%	50%	60%	70%
Oracle	73.62 ± 0.14	72.78 ± 0.15	71.78 ± 0.15	70.82 ± 0.15	69.27 ± 0.16	64.70 ± 0.17	60.59 ± 0.17	53.88 ± 0.18
Nearest $k = 1$	53.02 ± 0.19	49.04 ± 0.18	45.02 ± 0.18	40.87 ± 0.18	37.28 ± 0.17	33.13 ± 0.17	29.07 ± 0.16	24.64 ± 0.15
Nearest $k = 3$	53.79 ± 0.19	50.84 ± 0.18	47.24 ± 0.18	43.21 ± 0.17	38.58 ± 0.17	33.72 ± 0.16	29.00 ± 0.15	24.24 ± 0.13
Nearest $k = 5$	55.03 ± 0.20	53.08 ± 0.19	50.29 ± 0.19	46.50 ± 0.18	41.97 ± 0.17	36.51 ± 0.16	30.75 ± 0.15	25.26 ± 0.14
Linear Classifier	72.08 ± 0.14	68.56 ± 0.15	64.17 ± 0.16	58.90 ± 0.16	52.68 ± 0.16	45.43 ± 0.16	37.20 ± 0.15	29.18 ± 0.14
Matching Networks [54]	62.63 ± 0.19	60.81 ± 0.19	58.21 ± 0.19	54.79 ± 0.19	50.05 ± 0.19	43.47 ± 0.18	35.90 ± 0.17	28.70 ± 0.15
MAML [17]	64.37 ± 0.18	64.42 ± 0.18	55.27 ± 0.18	44.17 ± 0.18	44.10 ± 0.18	44.01 ± 0.18	32.03 ± 0.16	20.04 ± 0.13
Vanilla ProtoNet [46]	73.65 ± 0.14	71.80 ± 0.15	69.19 ± 0.15	65.28 ± 0.16	59.52 ± 0.17	51.42 ± 0.18	41.43 ± 0.18	32.29 ± 0.18
RNNP [35]	73.47 ± 0.14	71.80 ± 0.15	69.37 ± 0.16	65.88 ± 0.17	60.51 ± 0.18	52.25 ± 0.19	41.74 ± 0.19	32.47 ± 0.19
Median	73.54 ± 0.14	71.90 ± 0.15	69.30 ± 0.15	65.59 ± 0.16	59.88 ± 0.17	51.42 ± 0.18	41.13 ± 0.19	31.99 ± 0.18
Absolute $T = 10.0$	71.12 ± 0.15	69.58 ± 0.16	66.77 ± 0.17	62.27 ± 0.18	54.91 ± 0.20	45.13 ± 0.21	35.05 ± 0.20	28.20 ± 0.18
Absolute $T = 25.0$	73.10 ± 0.14	71.66 ± 0.15	69.13 ± 0.16	65.15 ± 0.17	58.63 ± 0.18	49.02 ± 0.19	38.40 ± 0.20	30.05 ± 0.18
Absolute $T = 50.0$	73.49 ± 0.14	71.88 ± 0.15	69.42 ± 0.16	65.52 ± 0.16	59.54 ± 0.18	50.65 ± 0.19	40.04 ± 0.19	31.34 ± 0.18
Euclidean $T = 10.0$	73.11 ± 0.15	71.60 ± 0.15	69.28 ± 0.16	65.57 ± 0.17	59.59 ± 0.18	50.45 ± 0.19	39.73 ± 0.19	30.88 ± 0.18
Euclidean $T = 25.0$	73.57 ± 0.14	71.98 ± 0.15	69.50 ± 0.16	65.78 ± 0.16	60.02 ± 0.18	51.59 ± 0.18	40.96 ± 0.19	31.98 ± 0.18
Euclidean $T = 50.0$	73.64 ± 0.14	71.96 ± 0.15	69.36 ± 0.16	65.68 ± 0.16	59.95 ± 0.17	51.58 ± 0.18	41.30 ± 0.19	32.18 ± 0.18
Cosine $T = 0.2$	73.62 ± 0.14	71.94 ± 0.15	69.44 ± 0.15	65.65 ± 0.16	59.91 ± 0.17	51.49 ± 0.18	41.14 ± 0.19	32.13 ± 0.18
Cosine $T = 0.5$	73.60 ± 0.14	71.85 ± 0.15	69.26 ± 0.15	65.46 ± 0.16	59.64 ± 0.17	51.50 ± 0.18	41.23 ± 0.19	32.18 ± 0.18
Cosine $T = 1.0$	73.57 ± 0.14	71.78 ± 0.15	69.13 ± 0.15	65.36 ± 0.16	59.62 ± 0.17	51.56 ± 0.18	41.44 ± 0.18	32.24 ± 0.18
Cosine $T = 2.0$	73.65 ± 0.14	71.83 ± 0.15	69.08 ± 0.16	65.25 ± 0.16	59.58 ± 0.17	51.26 ± 0.18	41.36 ± 0.18	32.10 ± 0.18
Cosine $T = 5.0$	73.55 ± 0.14	71.73 ± 0.15	69.06 ± 0.15	65.19 ± 0.16	59.42 ± 0.17	51.38 ± 0.18	41.31 ± 0.19	32.19 ± 0.18
TraNFS-2	72.80 ± 0.15	71.86 ± 0.15	70.54 ± 0.16	68.25 ± 0.17	64.29 ± 0.19	57.04 ± 0.21	45.84 ± 0.24	35.09 ± 0.23
TraNFS-3	73.17 ± 0.15	72.14 ± 0.15	70.71 ± 0.16	68.48 ± 0.17	64.59 ± 0.18	57.45 ± 0.21	45.80 ± 0.24	35.12 ± 0.23

C.1. 3-shot MiniImageNet

We show 5-way 3-shot performance on MiniImageNet with symmetric label swap (Table 6) and outlier (Table 7) noise. Note that we do not show results for paired label swap noise, as at 33.3% noise, paired label noise is identical to symmetric, and at 66.7%, the clean class is dominated by the noisy class.

We observe similar trends as in the 5-way 5-shot experiments reported in Tables 1, 2, and 3. The baseline methods suffer dramatically from replacing a clean sample in the support set with a single noisy sample, with ProtoNet [46] suffering almost a 14% drop in accuracy in the 33.3% sym-

metric label swap noise setting, as compared to the 5.71% drop in accuracy from removing a shot. Our proposed ProtoNet variants at various temperatures T all outperform vanilla ProtoNet. On the other hand, our TraNFS surpasses vanilla ProtoNet by 5.06% and impressively is only 3.05% short of the Oracle, despite not having knowledge of the noisy samples within the support set.

C.2. 10-shot MiniImageNet

We show 5-way 10-shot performance on MiniImageNet with symmetric label swap (Table 8), paired label swap (Table 9), and outlier (Table 10) noise. Note that we only show

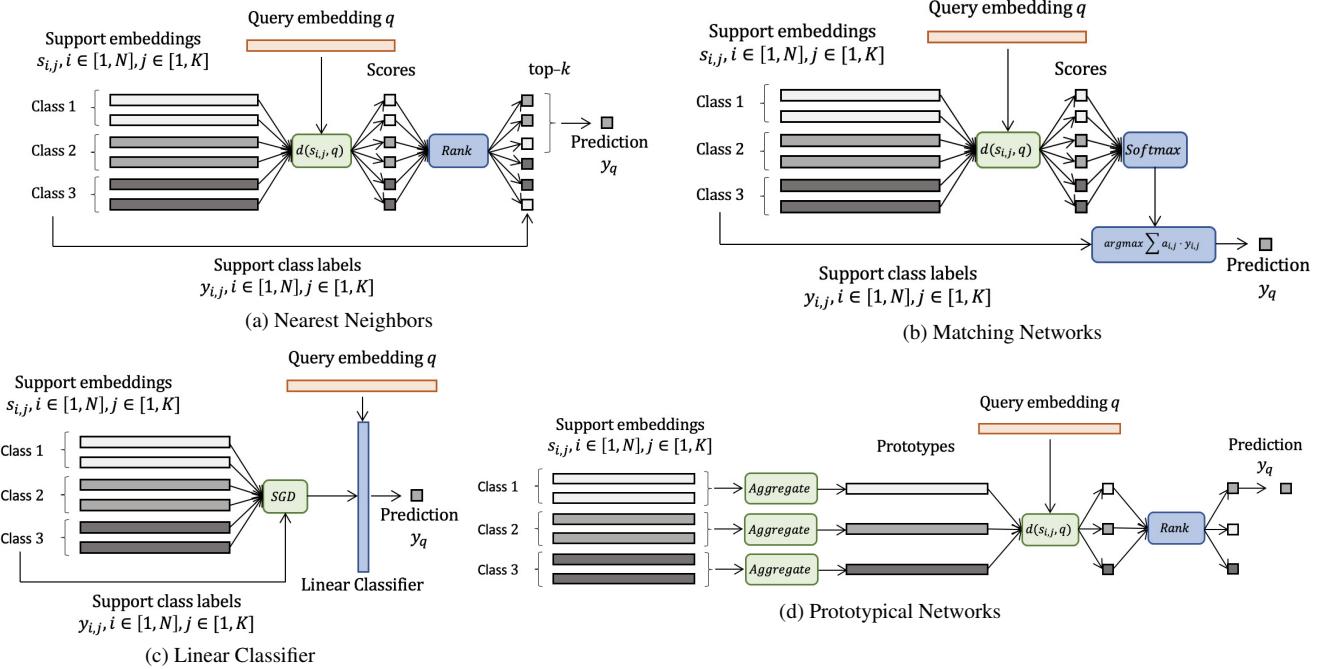


Figure 6. Visual overview of several of the few-shot method archetypes considered.

Table 9. Few-shot performance with paired label swap noise on 5-way 10-shot MiniImageNet [54].

Model \ Noise Proportion	20%	30%	40%
Oracle	71.78 ± 0.15	70.82 ± 0.15	69.27 ± 0.16
Nearset $k = 1$	44.85 ± 0.18	40.80 ± 0.18	36.58 ± 0.17
Nearset $k = 3$	46.96 ± 0.18	42.31 ± 0.17	37.32 ± 0.16
Nearset $k = 5$	49.88 ± 0.18	45.21 ± 0.17	39.47 ± 0.17
Linear Classifier	63.54 ± 0.16	56.70 ± 0.16	47.85 ± 0.16
Matching Networks [54]	57.74 ± 0.19	52.80 ± 0.18	45.37 ± 0.17
MAML [17]	55.05 ± 0.18	41.95 ± 0.18	41.83 ± 0.18
Vanilla ProtoNet [46]	68.34 ± 0.16	62.59 ± 0.16	52.73 ± 0.17
RNNP [35]	68.89 ± 0.16	63.86 ± 0.17	54.06 ± 0.18
Median	69.04 ± 0.15	63.50 ± 0.16	53.61 ± 0.17
Absolute $T = 50.0$	69.07 ± 0.16	63.62 ± 0.17	53.78 ± 0.18
Absolute $T = 25.0$	69.00 ± 0.16	63.75 ± 0.17	53.88 ± 0.18
Absolute $T = 10.0$	66.94 ± 0.17	61.82 ± 0.18	52.28 ± 0.20
Euclidean $T = 50.0$	68.86 ± 0.16	63.20 ± 0.17	53.24 ± 0.17
Euclidean $T = 25.0$	69.12 ± 0.16	63.48 ± 0.17	53.58 ± 0.17
Euclidean $T = 10.0$	68.91 ± 0.16	63.73 ± 0.17	53.81 ± 0.18
Cosine $T = 5.0$	68.50 ± 0.15	62.72 ± 0.16	52.76 ± 0.17
Cosine $T = 2.0$	68.42 ± 0.15	62.63 ± 0.16	52.87 ± 0.17
Cosine $T = 1.0$	68.48 ± 0.16	62.59 ± 0.17	52.86 ± 0.17
Cosine $T = 0.5$	68.58 ± 0.15	62.76 ± 0.16	52.90 ± 0.17
Cosine $T = 0.2$	68.82 ± 0.16	63.14 ± 0.17	53.27 ± 0.17
TraNFS-2	70.13 ± 0.16	66.20 ± 0.17	56.97 ± 0.20
TraNFS-3	70.38 ± 0.16	67.03 ± 0.18	58.94 ± 0.21

20%, 30%, and 40% noise proportion for paired label swap noise, as at 0% and 10%, paired label swapping is no different from symmetric swapping (Table 8) for 10 shots, and at 50% and above the noisy class would have either a share of or the outright majority.

Our proposed TraNFS shines with 10-shot tasks as well.

As in the 5-shot case, our method does especially well in moderate to high noise levels. In particular, we observe over 5% absolute improvement from TraNFS over vanilla ProtoNet at 40% and 50% symmetric label swap noise and an impressive 6.21% improvement for 40% paired label swap noise. TraNFS is also the best method for rejecting outlier noise as well.

D. Method descriptions

Fig. 6 shows a visual comparison of some of the baselines we compare against. We discuss implementation details below.

Oracle. When noise appears in a support set, accuracy of the few-shot model is reduced for two reasons: (1) mislabeled samples provide the model with misleading information about the class, and (2) clean samples that would have otherwise been informative were removed from the support set. FSL performance can be heavily influenced by the number of shots, especially in the low-data regime, so we find it important to separate out the aforementioned two sources of performance degradation. For this purpose, we include in our results tables an *Oracle* model consisting of a vanilla ProtoNet [46] with prototypes produced from only the correctly labeled samples in the support set. Note that the Oracle requires knowing the identities of the noisy samples, which cannot be reasonably expected in many real-world settings and is thus not a fair comparison with the other methods, but we include it to give a sense of constructive

Table 10. Few-shot performance with outlier noise on 5-way 10-shot MiniImageNet [54].

Model \ Noise Proportion	0%	10%	20%	30%	40%	50%	60%	70%
Oracle	73.62 ± 0.14	72.78 ± 0.15	71.78 ± 0.15	70.82 ± 0.15	69.27 ± 0.16	64.70 ± 0.17	60.59 ± 0.17	53.88 ± 0.18
Nearest $k = 1$	53.14 ± 0.19	50.61 ± 0.19	48.25 ± 0.18	45.62 ± 0.18	42.91 ± 0.18	39.99 ± 0.17	37.06 ± 0.17	33.66 ± 0.17
Nearest $k = 3$	53.55 ± 0.19	51.49 ± 0.18	49.16 ± 0.18	46.50 ± 0.18	43.69 ± 0.17	40.63 ± 0.17	37.07 ± 0.16	33.15 ± 0.15
Nearest $k = 5$	54.81 ± 0.20	53.31 ± 0.19	51.46 ± 0.19	49.18 ± 0.18	46.35 ± 0.18	43.13 ± 0.17	39.32 ± 0.16	35.05 ± 0.16
Linear Classifier	71.90 ± 0.15	69.62 ± 0.15	66.94 ± 0.16	63.70 ± 0.16	59.86 ± 0.16	55.31 ± 0.17	49.84 ± 0.17	43.42 ± 0.17
Matching Networks [54]	62.68 ± 0.19	61.37 ± 0.19	59.58 ± 0.19	57.52 ± 0.19	54.58 ± 0.19	51.12 ± 0.19	46.48 ± 0.19	40.68 ± 0.18
MAML [17]	64.30 ± 0.18	64.43 ± 0.18	58.82 ± 0.18	51.30 ± 0.19	51.37 ± 0.19	51.36 ± 0.19	42.05 ± 0.19	30.89 ± 0.18
Vanilla ProtoNet [46]	73.67 ± 0.14	72.27 ± 0.15	70.55 ± 0.15	68.08 ± 0.16	64.93 ± 0.16	60.66 ± 0.17	55.28 ± 0.18	47.94 ± 0.19
RNNP [35]	73.35 ± 0.14	71.92 ± 0.15	70.16 ± 0.15	67.97 ± 0.16	64.90 ± 0.17	60.81 ± 0.17	55.34 ± 0.18	48.07 ± 0.19
Median	73.69 ± 0.14	72.50 ± 0.15	70.78 ± 0.15	68.47 ± 0.15	65.26 ± 0.16	61.07 ± 0.17	55.46 ± 0.18	47.92 ± 0.19
Absolute $T = 50.0$	73.56 ± 0.14	72.44 ± 0.15	70.82 ± 0.15	68.60 ± 0.16	65.48 ± 0.16	61.33 ± 0.17	55.62 ± 0.18	48.19 ± 0.19
Absolute $T = 25.0$	73.26 ± 0.14	72.14 ± 0.15	70.65 ± 0.15	68.57 ± 0.16	65.53 ± 0.17	61.29 ± 0.17	55.45 ± 0.18	47.89 ± 0.19
Absolute $T = 10.0$	71.10 ± 0.15	69.96 ± 0.15	68.48 ± 0.16	66.29 ± 0.17	63.36 ± 0.17	58.83 ± 0.18	52.58 ± 0.19	44.80 ± 0.20
Euclidean $T = 50.0$	73.62 ± 0.14	72.39 ± 0.15	70.59 ± 0.15	68.28 ± 0.16	65.21 ± 0.16	60.94 ± 0.17	55.37 ± 0.18	48.04 ± 0.19
Euclidean $T = 25.0$	73.58 ± 0.14	72.36 ± 0.15	70.70 ± 0.15	68.40 ± 0.16	65.33 ± 0.16	61.08 ± 0.17	55.15 ± 0.18	47.79 ± 0.19
Euclidean $T = 10.0$	73.19 ± 0.15	72.03 ± 0.15	70.48 ± 0.16	68.21 ± 0.16	65.00 ± 0.17	60.50 ± 0.18	54.51 ± 0.19	46.58 ± 0.20
Cosine $T = 5.0$	73.57 ± 0.14	72.25 ± 0.15	70.44 ± 0.15	67.97 ± 0.16	64.77 ± 0.16	60.61 ± 0.17	55.14 ± 0.18	47.94 ± 0.19
Cosine $T = 2.0$	73.63 ± 0.14	72.28 ± 0.14	70.47 ± 0.15	68.10 ± 0.16	64.79 ± 0.16	60.60 ± 0.17	55.02 ± 0.18	48.03 ± 0.19
Cosine $T = 1.0$	73.46 ± 0.14	72.19 ± 0.15	70.33 ± 0.15	67.97 ± 0.16	64.80 ± 0.16	60.59 ± 0.17	55.09 ± 0.18	47.88 ± 0.19
Cosine $T = 0.5$	73.64 ± 0.14	72.30 ± 0.15	70.53 ± 0.15	68.13 ± 0.16	65.07 ± 0.16	60.73 ± 0.17	55.16 ± 0.18	48.13 ± 0.19
Cosine $T = 0.2$	73.55 ± 0.14	72.40 ± 0.15	70.61 ± 0.15	68.37 ± 0.16	65.26 ± 0.16	61.00 ± 0.17	55.34 ± 0.18	47.97 ± 0.19
TraNFS-2	72.43 ± 0.15	71.54 ± 0.16	70.24 ± 0.16	68.56 ± 0.17	65.93 ± 0.18	62.21 ± 0.20	56.98 ± 0.21	49.41 ± 0.22
TraNFS-3	72.91 ± 0.15	72.12 ± 0.15	70.92 ± 0.16	69.47 ± 0.16	67.14 ± 0.17	63.60 ± 0.19	58.68 ± 0.20	50.66 ± 0.22

information content still available in the support set after noise corruption.

Nearest Neighbors. In the context of FSL, nearest neighbors (Fig. 6a) is a simple, non-parametric classification technique which classifies query samples based on the labels of the k closest support samples in embedding space. Whichever class has the plurality among the k nearest neighbor support samples is the prediction, with ties broken uniformly at random among the tied classes. We report results for $k \in \{1, 3, 5\}$.

Linear Classifier. We train a single fully connected layer $\mathbb{R}^D \rightarrow \mathbb{R}^N$ on top of frozen convolutional features (Fig. 6c). For each episode, the parameters of the fully connected layer are learned with the AdamW [32] optimizer with weight decay 0.01, trained for 100 steps. Note that this approach resembles the Baseline method [11], with the primary difference being that we use the ProtoNet objective and episodic meta-training to learn the feature extractor \mathcal{F} , as opposed to the softmax cross entropy loss with batch learning on the base classes.

Matching Networks [54]. Matching networks (Fig. 6b) use an attention mechanism to compare the embedded query sample with embeddings of each of the support set samples, with the prediction being a linear combination of the support set labels based on the result of this attention. While this mechanism is trainable in a meta-learning setup, we found that we achieved better results than those reported in the literature by using a frozen convolutional feature extractor trained with the ProtoNet loss.

MAML [17]. Model-Agnostic Meta-Learning (MAML) seeks to learn a good initialization so that the model can be quickly adapted to new tasks, with this initialization learned through second-order gradients. As such, unlike the other methods we compare against, we do not use the weights of the same frozen 4-layer convolutional feature extractor for MAML. Instead, we use the Adam optimizer to train MAML with a meta-learning rate of 3×10^{-3} and inner loop learning rate of 1×10^{-2} , using 5 adaptation steps during meta-training and 10 steps during meta-test. We use the same random horizontal flips, resized crops, and color jitters for data augmentations as the rest of our experiments.

ProtoNet [46]. ProtoNet (Fig. 6d) was introduced in Sec. 3. We refer to the version of ProtoNet proposed by Snell *et al.* in [46] (using the mean of the support embeddings) as *Vanilla ProtoNet* to distinguish it from the median and similarity weighted variants of ProtoNet that we propose in Sec. 4.

Baseline++ [11]. Baseline++ was proposed as a simple alternative to recent few-shot methods. Rather than requiring relatively complex bi-level meta-training, [11] proposed simply pre-training a feature extractor with a standard supervised cross-entropy loss, freezing the feature extractor’s weights, and then fine-tuning a one-layer classifier just on top of the few examples in the novel class’s support set features. In particular, the Baseline++ method uses cosine similarity and a softmax for the classifier. Such an approach has been shown to be surprisingly competitive with popular few-shot approaches. We implement this cosine similarity

Table 11. **Temperature sweep for our ProtoNet variants: symmetric label swap noise.** 5-way 5-shot Acc. \pm 95% CI on MiniImageNet [54], TieredImageNet [44]. Best viewed in color.

Model \ Noise Proportion	0%		20%		40%		60%	
Absolute $T = 50.0$	68.18 \pm 0.16	71.24 \pm 0.18	62.98 \pm 0.17	66.56 \pm 0.20	51.68 \pm 0.19	54.97 \pm 0.21	39.24 \pm 0.20	41.59 \pm 0.21
Absolute $T = 25.0$	68.24 \pm 0.16	71.27 \pm 0.18	63.46 \pm 0.17	66.87 \pm 0.20	52.06 \pm 0.20	55.26 \pm 0.22	39.78 \pm 0.20	42.54 \pm 0.22
Absolute $T = 10.0$	67.15 \pm 0.17	70.15 \pm 0.19	62.96 \pm 0.18	66.10 \pm 0.20	52.08 \pm 0.20	55.08 \pm 0.23	39.92 \pm 0.21	42.49 \pm 0.23
Absolute $T = 5.0$	63.89 \pm 0.17	66.56 \pm 0.19	59.63 \pm 0.18	62.67 \pm 0.21	51.30 \pm 0.20	53.83 \pm 0.22	37.99 \pm 0.21	39.91 \pm 0.23
Absolute $T = 1.0$	50.26 \pm 0.20	51.39 \pm 0.22	47.04 \pm 0.20	48.40 \pm 0.23	40.40 \pm 0.21	41.45 \pm 0.23	31.03 \pm 0.20	31.75 \pm 0.21
Euclidean $T = 50.0$	68.31 \pm 0.16	71.31 \pm 0.18	62.78 \pm 0.17	66.36 \pm 0.19	51.86 \pm 0.19	55.19 \pm 0.21	38.90 \pm 0.20	41.19 \pm 0.21
Euclidean $T = 25.0$	68.32 \pm 0.16	71.48 \pm 0.18	63.02 \pm 0.17	66.69 \pm 0.19	52.09 \pm 0.19	55.62 \pm 0.21	39.33 \pm 0.20	41.75 \pm 0.21
Euclidean $T = 10.0$	68.23 \pm 0.16	71.18 \pm 0.19	63.46 \pm 0.17	67.04 \pm 0.20	52.24 \pm 0.20	55.78 \pm 0.22	39.87 \pm 0.20	42.53 \pm 0.22
Euclidean $T = 5.0$	67.53 \pm 0.16	70.54 \pm 0.18	63.00 \pm 0.18	66.56 \pm 0.20	53.79 \pm 0.20	57.37 \pm 0.22	39.63 \pm 0.21	42.31 \pm 0.22
Euclidean $T = 1.0$	56.75 \pm 0.19	59.17 \pm 0.21	52.31 \pm 0.19	54.82 \pm 0.22	44.06 \pm 0.20	46.09 \pm 0.23	32.88 \pm 0.20	33.99 \pm 0.21
Cosine $T = 10.0$	68.24 \pm 0.16	71.27 \pm 0.18	62.47 \pm 0.17	66.16 \pm 0.19	51.41 \pm 0.19	54.96 \pm 0.21	38.38 \pm 0.19	40.74 \pm 0.21
Cosine $T = 5.0$	68.31 \pm 0.16	71.16 \pm 0.18	62.51 \pm 0.17	65.99 \pm 0.20	51.51 \pm 0.19	54.78 \pm 0.21	38.55 \pm 0.19	40.81 \pm 0.21
Cosine $T = 2.0$	68.28 \pm 0.16	71.22 \pm 0.18	62.57 \pm 0.17	66.24 \pm 0.19	51.59 \pm 0.19	55.06 \pm 0.21	38.71 \pm 0.19	40.99 \pm 0.21
Cosine $T = 1.0$	68.21 \pm 0.16	71.21 \pm 0.18	62.70 \pm 0.17	66.47 \pm 0.19	51.72 \pm 0.19	55.27 \pm 0.21	38.92 \pm 0.19	41.32 \pm 0.21
Cosine $T = 0.5$	68.42 \pm 0.16	71.31 \pm 0.18	63.13 \pm 0.18	66.81 \pm 0.20	52.08 \pm 0.19	55.60 \pm 0.22	39.36 \pm 0.20	42.14 \pm 0.22
Cosine $T = 0.2$	68.20 \pm 0.16	70.59 \pm 0.18	63.46 \pm 0.17	66.62 \pm 0.20	52.42 \pm 0.20	55.78 \pm 0.22	39.90 \pm 0.20	42.56 \pm 0.22
Cosine $T = 0.1$	67.52 \pm 0.16	69.30 \pm 0.19	63.07 \pm 0.18	65.25 \pm 0.20	52.22 \pm 0.20	54.24 \pm 0.23	39.85 \pm 0.21	41.79 \pm 0.23

Table 12. **Temperature sweep for our ProtoNet variants: paired label swap noise.** 5-way 5-shot Acc. \pm 95% CI on MiniImageNet [54], TieredImageNet [44]. Best viewed in color.

Model \ Noise Proportion	40%	
Absolute $T = 50.0$	48.64 \pm 0.19	51.83 \pm 0.21
Absolute $T = 25.0$	49.38 \pm 0.20	52.40 \pm 0.22
Absolute $T = 10.0$	49.56 \pm 0.20	52.54 \pm 0.23
Absolute $T = 5.0$	47.18 \pm 0.21	49.42 \pm 0.23
Absolute $T = 1.0$	37.85 \pm 0.21	38.47 \pm 0.23
Euclidean $T = 50.0$	48.43 \pm 0.19	51.39 \pm 0.21
Euclidean $T = 25.0$	48.67 \pm 0.19	51.90 \pm 0.21
Euclidean $T = 10.0$	49.37 \pm 0.19	52.55 \pm 0.22
Euclidean $T = 5.0$	49.75 \pm 0.20	52.57 \pm 0.22
Euclidean $T = 1.0$	41.30 \pm 0.21	42.92 \pm 0.22
Cosine $T = 10.0$	47.75 \pm 0.19	50.95 \pm 0.21
Cosine $T = 5.0$	48.03 \pm 0.19	51.17 \pm 0.21
Cosine $T = 2.0$	48.03 \pm 0.19	51.19 \pm 0.21
Cosine $T = 1.0$	48.53 \pm 0.19	51.71 \pm 0.21
Cosine $T = 0.5$	48.90 \pm 0.19	52.14 \pm 0.21
Cosine $T = 0.2$	49.40 \pm 0.19	52.72 \pm 0.22
Cosine $T = 0.1$	49.71 \pm 0.20	51.96 \pm 0.23

classifier in our framework, with the primary difference being that we use a feature extractor trained with the ProtoNet loss instead of a cross-entropy loss, in order to compare the classifier design on even terms. Note that [11] also proposed a simpler approach using a standard linear layer instead of cosine distance, which they referred to as Baseline; other than the training objective of the fixed feature extractor, the Baseline method is equivalent to our Linear Classifier baseline.

NegMargin [30]. Taking insights from the metric learning literature, [30] suggests that discriminability shortcomings of the softmax loss can be mitigated by learning with a margin. Surprisingly, NegMargin found that positive margins underperform in open-set few-shot classification scenarios, while negative margins can lead to significant improvements in performance due to improved transferabil-

ity. To perform few-shot classification, NegMargin takes a similar approach to [11]—first pre-training and then freezing the feature extractor, followed by fine-tuning of a classifier for the novel support set—with the primary difference being the substitution of the standard softmax with the negative margin softmax loss during pre-training. As such, unlike the other methods we compare against, we do not use the weights of the same frozen 4-layer convolutional feature extractor for NegMargin. We use the official NegMargin codebase,³ modifying their code to inject artificial noisy labels into support sets during meta-test evaluation.

RNNP [35]. Robust Nearest Neighbor Prototype (RNNP) creates hybrid examples by interpolating between samples within each support set, somewhat similarly to mixup. Using ProtoNet prototypes of the original support embeddings as initialization for the class centers, k -means is then used to refine the prototypes in an unsupervised manner. We reproduce RNNP, using the suggested $K - 1$ hybrids per support sample and mixing ratio of 0.8 when producing hybrids.

E. Additional ablation studies

Feature extractor training objective. We consider the performance of few-shot learning methods within the context of support set noise primarily with a frozen feature extractor, as is common practice in many previous few-shot works [11, 30, 46]. This allows us to isolate our comparison to the method, as opposed to the learned features. Nonetheless, the learned features have an impact on model performance. We compare the performance of 4-layer convolutional neural networks feature extractors [54] pre-trained with the ProtoNet [46] and NegMargin [30] objectives, observing $\{69.66 \pm 0.16, 59.88 \pm 0.18, 47.53 \pm 0.21\}$.

³<https://github.com/b10/negative-margin.few-shot>

Table 13. **Temperature sweep for our ProtoNet variants: outlier noise.** 5-way 5-shot Acc. \pm 95% CI on MiniImageNet [54], TieredImageNet [44]. Best viewed in color.

Model \ Noise Proportion	0%		20%		40%		60%	
Absolute $T = 50.0$	68.41 \pm 0.16	71.42 \pm 0.19	64.62 \pm 0.17	67.96 \pm 0.19	58.08 \pm 0.19	61.68 \pm 0.21	47.33 \pm 0.20	50.71 \pm 0.22
Absolute $T = 25.0$	68.13 \pm 0.16	71.17 \pm 0.18	64.69 \pm 0.17	68.00 \pm 0.19	58.30 \pm 0.18	61.98 \pm 0.21	47.39 \pm 0.20	50.59 \pm 0.22
Absolute $T = 10.0$	67.18 \pm 0.16	70.10 \pm 0.19	64.14 \pm 0.17	67.29 \pm 0.20	58.12 \pm 0.19	61.65 \pm 0.21	47.02 \pm 0.21	49.68 \pm 0.22
Absolute $T = 5.0$	63.97 \pm 0.17	66.78 \pm 0.19	60.96 \pm 0.18	63.88 \pm 0.20	55.24 \pm 0.19	58.17 \pm 0.21	44.28 \pm 0.21	46.50 \pm 0.22
Absolute $T = 1.0$	50.02 \pm 0.19	51.77 \pm 0.22	47.71 \pm 0.20	49.01 \pm 0.22	42.90 \pm 0.20	44.45 \pm 0.23	34.52 \pm 0.20	35.08 \pm 0.21
Euclidean $T = 50.0$	68.31 \pm 0.16	71.14 \pm 0.18	64.25 \pm 0.17	67.53 \pm 0.19	57.43 \pm 0.18	60.95 \pm 0.21	47.06 \pm 0.20	50.34 \pm 0.21
Euclidean $T = 25.0$	68.51 \pm 0.16	71.28 \pm 0.18	64.57 \pm 0.17	67.89 \pm 0.19	58.01 \pm 0.18	61.61 \pm 0.20	47.25 \pm 0.20	50.49 \pm 0.21
Euclidean $T = 10.0$	68.19 \pm 0.16	71.20 \pm 0.18	64.55 \pm 0.17	68.02 \pm 0.19	58.17 \pm 0.19	62.00 \pm 0.21	47.24 \pm 0.20	50.86 \pm 0.22
Euclidean $T = 5.0$	67.58 \pm 0.16	70.45 \pm 0.18	64.25 \pm 0.17	67.55 \pm 0.19	57.82 \pm 0.19	61.69 \pm 0.21	46.34 \pm 0.20	50.21 \pm 0.22
Euclidean $T = 1.0$	56.94 \pm 0.18	59.04 \pm 0.21	53.59 \pm 0.19	55.57 \pm 0.22	47.23 \pm 0.20	49.63 \pm 0.22	37.32 \pm 0.20	39.37 \pm 0.22
Cosine $T = 10.0$	68.41 \pm 0.16	71.20 \pm 0.18	64.19 \pm 0.17	67.48 \pm 0.19	57.33 \pm 0.18	60.87 \pm 0.21	47.02 \pm 0.20	50.12 \pm 0.21
Cosine $T = 5.0$	68.29 \pm 0.16	71.28 \pm 0.19	64.04 \pm 0.17	67.46 \pm 0.20	57.30 \pm 0.18	61.10 \pm 0.21	47.08 \pm 0.20	50.27 \pm 0.21
Cosine $T = 2.0$	68.29 \pm 0.16	71.20 \pm 0.18	64.13 \pm 0.17	67.53 \pm 0.19	57.39 \pm 0.18	61.07 \pm 0.20	46.97 \pm 0.20	50.23 \pm 0.21
Cosine $T = 1.0$	68.30 \pm 0.16	71.54 \pm 0.18	64.23 \pm 0.17	68.07 \pm 0.19	57.51 \pm 0.18	61.82 \pm 0.20	46.89 \pm 0.20	50.82 \pm 0.21
Cosine $T = 0.5$	68.35 \pm 0.16	71.38 \pm 0.18	64.51 \pm 0.17	68.16 \pm 0.19	57.97 \pm 0.18	62.13 \pm 0.20	47.28 \pm 0.20	51.14 \pm 0.22
Cosine $T = 0.2$	68.20 \pm 0.16	70.79 \pm 0.18	64.78 \pm 0.17	67.94 \pm 0.19	58.36 \pm 0.18	62.37 \pm 0.21	47.34 \pm 0.20	51.12 \pm 0.22
Cosine $T = 0.1$	67.82 \pm 0.16	69.33 \pm 0.19	64.49 \pm 0.17	66.55 \pm 0.20	58.42 \pm 0.19	61.21 \pm 0.21	46.90 \pm 0.21	50.00 \pm 0.22

Table 14. **Ablation study: Clean Prototype Loss** for a 3-layer TraNFS trained on 5-way 5-shot MiniImageNet [54].

λ_c	0%	20%	40%	60%
0.0	63.77 \pm 0.18	60.67 \pm 0.19	53.14 \pm 0.22	39.75 \pm 0.23
0.1	65.68 \pm 0.18	61.94 \pm 0.19	53.45 \pm 0.22	39.20 \pm 0.24
0.5	68.11 \pm 0.17	64.56 \pm 0.18	56.47 \pm 0.21	41.94 \pm 0.24
1.0	68.80 \pm 0.16	65.10 \pm 0.18	57.26 \pm 0.21	42.82 \pm 0.24
5.0	68.53 \pm 0.17	65.08 \pm 0.18	56.65 \pm 0.21	42.60 \pm 0.24
10.0	68.76 \pm 0.17	64.87 \pm 0.18	56.76 \pm 0.21	42.17 \pm 0.24

0.18, $35.67 \pm 0.17\}$ on 5-way 5-shot MiniImageNet [54] with $\{0\%, 20\%, 40\%, 60\%\}$ symmetric label swap noise. As reported in the literature, NegMargin outperforms the ProtoNet pre-trained feature extractor when there is no support set noise during meta-test. On the other hand, NegMargin sees a steeper decline in performance with increasing noise levels. We thus focus on the ProtoNet pre-trained feature extractor for our primary experiments. We leave further investigation into this phenomenon and the performance of other feature extractor pre-training objectives on noisy few-shot learning to future work.

Proposed ProtoNet variants: Temperature settings. As explained in Sec. 4.2, the temperature T controls the diffuseness of the softmax for similarity weighted prototypes. The setting of T results in a trade-off between emphasizing more shots versus noise rejection capability and thus can have an impact on performance. We show performance of similarity weighted prototypes with absolute distance, squared euclidean distance, and cosine similarity measure on MiniImageNet and TieredImageNet at varying noise levels with symmetric label swap noise, paired label swap noise, and outlier noise in Tables 11, 12, and 13, respectively. Note that differences in scale of T for Absolute and Squared Euclidean distances versus cosine similarity is due to their scale: cosine similarity is within $[-1, 1]$, while

Table 15. **Ablation study: Binary Classification Loss** for a 3-layer TraNFS trained on 5-way 5-shot MiniImageNet [54].

λ_b	0%	20%	40%	60%
0.0	68.74 \pm 0.17	64.97 \pm 0.18	56.29 \pm 0.21	41.88 \pm 0.23
0.1	68.73 \pm 0.17	65.04 \pm 0.18	56.57 \pm 0.21	42.23 \pm 0.24
0.5	68.53 \pm 0.17	65.08 \pm 0.18	56.65 \pm 0.21	42.60 \pm 0.24
1.0	68.74 \pm 0.17	64.81 \pm 0.18	56.44 \pm 0.21	42.26 \pm 0.24
5.0	68.75 \pm 0.17	65.06 \pm 0.18	56.71 \pm 0.21	42.42 \pm 0.24

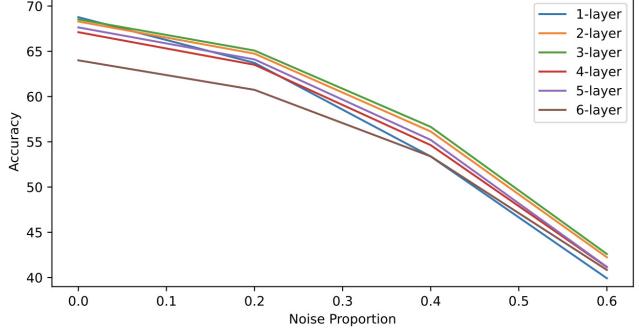


Figure 7. Sweep of number of transformer layers for 5-way 5-shot MiniImageNet [54] with symmetric label swap noise.

the two distances depend on the feature dimensionality and scale.

TraNFS: Clean prototype loss. We run a hyperparameter sweep for the loss weight term λ_c , which controls the weight of the clean prototype loss (Eq. (15)). Results are reported in Table 14. We observe that the clean prototype loss is indeed helpful for encouraging the transformer to learn how to reject noisy samples, with a range of values of λ_c that work well.

TraNFS: Binary outlier detection. To test the effectiveness of the binary outlier classifier loss (Eq. (16)), we run a

Table 16. **Ablation study: choice of embedding for `CLS` tokens** for a 3-layer TraNFS trained on 5-way 5-shot MiniImageNet [54] with symmetric label swap noise.

CLS Token + POS Token	0%	20%	40%	60%
Prototype + Learnable	68.15 ± 0.16	64.68 ± 0.18	55.04 ± 0.21	41.12 ± 0.22
Learnable + Learnable	67.74 ± 0.17	64.28 ± 0.18	55.46 ± 0.22	41.42 ± 0.24
Random Constant + Random Constant	66.95 ± 0.17	63.34 ± 0.19	54.55 ± 0.22	40.87 ± 0.24
Random Constant + Learnable	68.53 ± 0.17	65.08 ± 0.18	56.65 ± 0.21	42.60 ± 0.24

hyperparameter sweep for the loss weight term λ_b , reporting results in Table 15. We find that binary outlier classifier is indeed effective, with relatively low sensitivity to the setting of λ_b . Thus, we set λ_b to be 0.5 throughout our other experiments.

TraNFS: `CLS` and `POS` token embeddings. There are several options for the embeddings, corresponding to the `CLS` and `POS` tokens. In Table 16, we meta-train a 3-layer TraNFS model on 5-shot 5-way MiniImageNet with symmetric label swap noise. Each class’s `CLS` token is set using one of three options: class prototypes averaged from the convolutional embeddings, a learnable parameter, and a random constant.

While we expected the ProtoNet-style prototypes to help kick-start the transformer’s comparison mechanism, we were surprised to instead observe that they underperform other choices for the `CLS` embeddings. After visualizing the learning curves, we observe that using prototypes as the `CLS` embeddings results in a difficult-to-escape local minimum; we hypothesize this may be the model having minimal incentive to learn anything beyond the provided prototype. We also find that learnable `CLS` embeddings are not particularly effective: due to the random identity and shuffling of class orders between tasks, each `CLS` embedding lacks any semantic meaning beyond corresponding to a particular `POS` token’s support samples; thus trying to learn some discriminative value does not transfer between tasks and is ultimately unhelpful. As a result, it appears that a random constant value for each `CLS` token is sufficient for the transformer. For the `POS` positional encodings, however, learnable embeddings seem to work best.

TraNFS: Number of layers. Fig. 7 reports results for a sweep over the number of transformer layers in TraNFS. Matching intuition, we find that one layer is insufficient for surpassing the mean (ProtoNet) baseline. Different classes for each N -way episode mean the `CLS` embedding do not generalize across tasks. Without prior information of what each class c is in an episode, the transformer needs at least one layer to form such a concept for each position before comparisons can be made to identify samples that do not belong. Training with too many layers, however, seems to occasionally be unstable and tends to produce slightly inferior results, perhaps due to too much overparameterization and overfitting. We find two or three layers tend to perform best and thus report most of our results as such. More

layers in the transformer of course increases computational costs, but techniques such as knowledge distillation [21, 29] can be used to reduce model size while minimizing performance loss.