Proceedings of Machine Learning Research 140:45–57, 2021 AAAI Workshop on Meta-Learning and MetaDL Challenge

Semi-Supervised Few-Shot Learning with Prototypical Random Walks

Ahmed Ayyad A.3AYAD@GMAIL.COM
Technical University of Munich, Munich, Germany
Yuchen Li YUCHEN.LI@KAUST.EDU.SA
King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
Raden Muaz RADEN.M.MUAZ@GMAIL.COM
King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
Shadi Albarqouni* SHADI.ALBARQOUNI@HELMHOLTZ-MUENCHEN.DE
Technical University of Munich, Munich, Germany & Helmholtz AI, Helmholtz Center Munich, Neuherberg, Germany
Mohamed Elhoseiny* MOHAMED.ELHOSEINY@KAUST.EDU.SA
King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Editors: Isabelle Guyon, Jan N. van Rijn, Sébastien Treguer, Joaquin Vanschoren

Abstract

Recent progress has shown that few-shot learning can be improved with access to unlabelled data, known as semi-supervised few-shot learning(SS-FSL). We introduce an SS-FSL approach, dubbed as Prototypical Random Walk Networks(PRWN), built on top of Prototypical Networks (PN). We develop a random walk semi-supervised loss that enables the network to learn representations that are compact and well-separated. Our work is related to the very recent development of graph-based approaches for few-shot learning. However, we show that compact and well-separated class representations can be achieved by modeling our prototypical random walk notion without needing additional graph-NN parameters or requiring a transductive setting where a collective test set is provided. Our model outperforms baselines in most benchmarks with significant improvements in some cases. Our model, trained with 40% of the data as labeled, compares competitively against fully supervised prototypical networks, trained on 100% of the labels, even outperforming it in the 1-shot mini-Imagenet case with 50.89% to 49.4% accuracy. We also show that our loss is resistant to distractors, unlabeled data that does not belong to any of the training classes, and hence reflecting robustness to labeled/unlabeled class distribution mismatch. The associated GitHub page can be found at https://prototypical-random-walk.github.io.

1. Introduction

Few-shot learning is an artificial learning skill of rapidly generalizing from limited supervisory data (few labeled examples), typically without the use of any unlabeled data (Koch et al., 2015; Miller et al., 2000; Lake et al., 2011). Our work is at the intersection between few-shot learning and semi-supervised learning, where we augment the capability of few-shot artificial learners with a learning signal derived from unlabeled data.

* Shared senior authorship

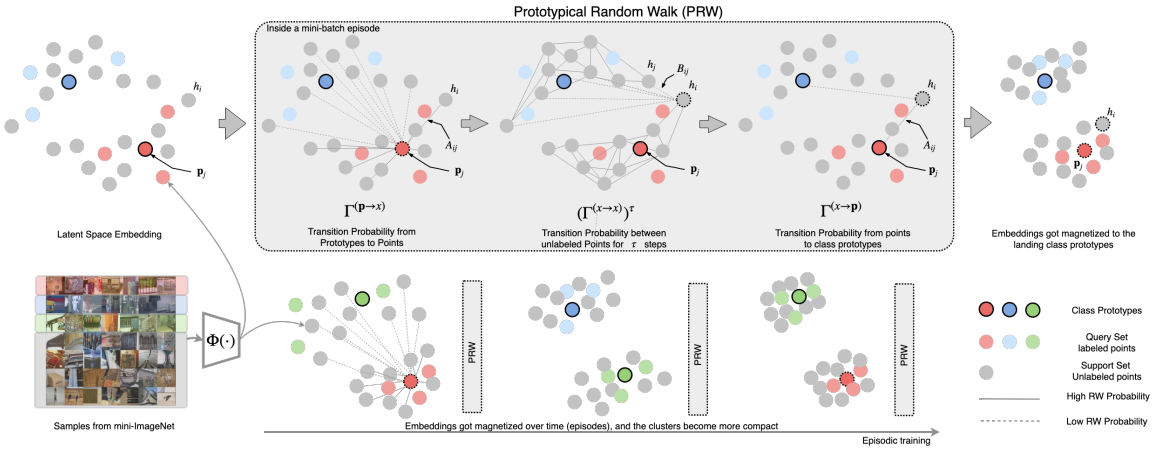


Figure 1: Our PRW aims at maximizing the probability of a random walk begins at the class prototype $\mathbf{p}_j$, taking τ steps among the unlabeled data, before it lands to the same class prototype. This results in a more discriminative representation, where the embedding of the unlabeled data of a particular class got magnetized to its corresponding class prototype, denoted as *prototypical magnetization*.

Semi-supervised Few-shot Learning (SSFSL): Few-shot learning methods typically applied the supervised learning setup (e.g., (Vinyals et al., 2016; Ravi and Larochelle, 2017b; Snell et al., 2017)), recently, Ren et al. (2018) and Zhang et al. (2018) developed Semi-super- vised few-shot learning approaches that can leverage additional unlabeled data. The machinery of both approaches adopts a meta-learning episodic training procedure with integrated learning signals from unlabeled data. Ren et al. (2018) built on the top of prototypical networks (PN) (Snell et al., 2017) so better class prototypes can be learned with the help of the unlabeled data. Zhang et al. (2018) proposed a GAN-based approach, Meta-GAN, that helps to make it easier for FSL models to learn better decision boundaries between different classes.

In this work, we propose Prototypical Random Walks (PRW) as an effective graph-based learning signal derived from unlabeled data. Our approach improves few-shot learning models by a prototypical random walk through the embeddings of unlabeled data starting from each class prototype passing through unlabeled data in the embedding space and encourages returning to the same prototype at the end of the prototypical walk (*cf.* Fig. 1). This PRW learning signal promotes a latent space where points of the same class are compactly clustered around their prototype while being well isolated from other prototypes. We sometimes refer to this discriminative attraction to class prototypes as *prototypical magnetization*.

Since the PRW loss is computed over a similarity graph involving all the prototypes and unlabeled points in the episode, it takes a global view of the data manifold. Due to the promoted *prototypical magnetization* property, this global view enables more efficient learning of discriminative embeddings from few examples, which is the key challenge in few-shot learning. In contrast, there are local SSL losses, where the loss is de-

fined over each point individually, most notable of those approaches is the state-of-the-art Virtual Adversarial Training (VAT) by Miyato et al. (2018). We show that in the FSL setting, our global consistency guided by our prototypical random walk loss adds a learning value compared to local consistency losses as in VAT (Miyato et al., 2018).

Contribution. We propose Prototypical Random Walk Networks (PRWN) where we promote *prototypical magnetization* of the learning representation. We demonstrate the effectiveness of PRWN on popular few-shot image classification benchmarks. We also show that our model trained with a fraction of the labels is competitive with PN trained with all the labels. Moreover, we demonstrate that our loss is robust to "distractor" points which could accompany the unlabeled data yet not belong to any of the training classes of the episode.

2. Approach

We build our approach on top of Prototypical Networks (PN) (Snell et al., 2017) and augment it with a novel random walk loss leveraging the unlabeled data during the *meta-training* phase. The key message of our work is that more discriminative few-shot representations can be learned through training with prototypical random walks. We maximize the probability of a random walk that starts from a class of prototypes and walk through the embeddings of unlabeled points to land back to the same prototype; see Fig. 1. Our random walk loss enforces the global consistency where the overall structure of the manifold is considered. In this section, we detail the problem definition and our loss.

2.1. Problem Set-up

The few-shot learning problem may be formulated as training over the distribution of classification tasks $\mathcal{P}_{train}(\mathcal{T})$, in order to generalize to a related distribution of tasks $\mathcal{P}_{test}(\mathcal{T})$ at test time. This setting entails two levels of learning; *meta-training* is learning the shared model parameters(meta-parameters) to be used on future tasks, *adaptation* is the learning done within each task. Meta-training can be seen as the outer training loop, while adaptation being the inner loop.

Concretely, for N_s-shot N_c-way FSL, each task is an episode with a support set $\mathcal{S}$ containing N_s labeled examples from each of N_c classes, and a query set $\mathcal{Q}$ of points to be classified into the N_c episode classes. The support set is used for adaptation, then the query set is used to evaluate our performance on the task and compute a loss for meta-training.

To run a standard FSL experiment, we split our datasets such that each class is present exclusively in one of our train/val/test splits. To generate a training episode, we sample N_c training classes from the train split and sample N_s samples from each class for the support set. Then we sample N_q images from the same classes for the query set. Validation and test episodes are sampled analogously from their respective splits.

Following the SS-FSL setup in (Ren et al., 2018; Zhang et al., 2018; Liu et al., 2019), we split our training dataset into labeled/ unlabeled; let $\mathcal{D}_{k,L}$ denote all labeled points $x \in class(k)$, and $\mathcal{D}_{k,U}$ be all unlabeled points $x \in class(k)$. Analogous notation holds for our support and query set, $\mathcal{S}$ and $\mathcal{Q}$. To set up a semi-supervised episode, we simply need to add some unlabeled data to the support set. For every class c sampled for the episode, we sample N_u samples from $\mathcal{D}_{c,U}$ and add them to $\mathcal{S}$. In order to make the setting more realistic and challenging, we also test our model with the addition of *distractor* data. Those are unlabeled points added to the support set, but not belonging to the episode classes. We simply sample N_d ad-

ditional classes, and sample N_u points from each class to add to the support set. We present pseudo-code for episode construction in the supp. materials.

It is worth mentioning that the unlabeled data may be present at either train or test time, or both. At training time, we want to use the unlabeled data for meta-training i.e. learning better model parameters. For unlabeled data at test time, we want to use it for better adaptation, i.e. performing better classification on the episode's query set. Our loss operates on the meta-training level, to leverage unlabeled data for learning better meta-parameters. However, we also present a version of our model capable of using unlabeled data for adaptation, by using the semi-supervised inference from Ren et al. (2018) with our trained models.

Prototypical Networks. Prototypical networks (Snell et al., 2017) aim to train a neural network as an embedding function mapping from input space to a latent space where points of the same class tend to cluster. The embedding function $\Phi(\cdot)$ is used to compute a *prototype* for each class, by averaging the embeddings of all points in the support belonging to that class:

$$\mathbf{p}_c = \frac{1}{|\mathcal{S}_{c,L}|} \sum_{x_i \in \mathcal{S}_{c,L}} \Phi(x_i; \theta) \tag{1}$$

where $\mathbf{p}_c$ is the prototype for our c-th class, and θ represents our meta-parameters. Once prototypes of all classes are obtained, query points are also embedded to the same space, and then classified based on their distances to the prototypes, via a softmax function. For a point x_i with an embedding $h_i = \Phi(x_i; \theta)$, the probability of belonging to class c is computed by

$$\begin{aligned} z_{i,c} = p(y_c|x_i) &= \frac{\exp\left(-d(h_i, \mathbf{p}_c)\right)}{\sum_{j=1}^{N_c} \exp\left(-d(h_i, \mathbf{p}_j)\right)}, \\ \tilde{\mathbf{p}}_c &= \frac{\sum_{x_i \in S_U \cup S_L} h_i \cdot z_{i,c}}{\sum_{i=1}^{N} z_{i,c}}. \end{aligned} \tag{2}$$

where $d(\cdot, \cdot)$ is the Euclidean distance. In the semi-supervised variant (Ren et al., 2018), PN use the unlabeled data to *refine* the class prototypes. This is achieved via a soft K-means step. First, the class probabilities for the unlabeled data $z_{i,c}$ are computed as in Eq.2, and the labeled points have a hard assignment, i.e. $z_{i,c}$ is 1 if $x_i \in class(c)$ and 0 otherwise. Then the updated prototype $\tilde{\mathbf{p}}_c$ is computed as the weighted average of the points assigned to it; see Eq. 2. We can see this as a task *adaptation* step, which does not directly propagate any learning signal from the unlabeled points to our model parameters θ. In fact, it might be used only at the inference time, and results from Ren et al. (2018) show that it provides a significant improvement when used as such. When used during *meta-training* by updating the model parameters from the unlabeled data, the performance improves only marginally (i.e., from 49.98% to 50.09% on mini-imagenet (Vinyals et al., 2016)). While this approach is powerful as the *adaptation* step, it fails to fully exploit the unlabeled data during *meta-training*. *SS-FSL with adaption at test time.* Our approach also allows using the former K-means refinement step at inference time, analogous to the 'Semi-supervised inference' model from Ren et al. (2018). Orthogonal to Ren et al. (2018), our approach can be thought of as a meta-training regularizer that brings discriminative global characteristics to the learning representation efficiently.

2.2. Prototypical Random Walk

Given the class prototypes $\mathbf{p}_c$, computed using the labeled data in the support set $\mathcal{S}_L$, and the embeddings h_i of unlabeled support set $\mathcal{S}_U$, we construct a similarity graph between the unlabeled points' embeddings and the prototypes. Our goal is to have points of the same class form a compact cluster in latent space, well separated from other classes.

Our Prototypical Random Walk(PRW) loss aims to aid this by compactly attracting the unlabeled embeddings around the class prototypes, promoting well-separation(*cf.* Fig. 1).

This notion is translated into the idea that a random walker over the similarity graph rarely crosses class decision boundaries. Here, we do not know the labels for our points or the right decision boundaries, so we cannot optimize for this directly. We basically imagine our walker starting at a prototype, taking a step to an unlabeled point, and then stepping back to a prototype. The objective is to increase the probability that the walker returns to the same prototype it started from; we will refer to this probability as the *landing probability*. Additionally, we let our walker taking some steps between the unlabeled points, before taking a step back to a prototype.

Concretely, for an episode with N class prototypes, and M unlabeled points overall, let $A \in \mathbb{R}^{M\times N}$ be the similarity matrix, such that each row contains the negative Euclidean distances between the embedding of an unlabeled point an d the class prototypes. Similarly, we compute the similarity matrix between the unlabeled points $B \in \mathbb{R}^{M\times M}$. Mathematically speaking, $A_{i,j} = -\|h_i - \mathbf{p}_j\|^2, B_{i,j} = -\|h_i - h_j\|^2$ where $h_i = \Phi(x_i)$ is the embedding of the i-th unlabeled sample, and $\mathbf{p}_j$ is the j-th class prototype. The diagonal entries $B_{i,i}$ are set to a small enough number to avoid self-loop.

Transition probability matrices for our random walker are calculated by taking a softmax over the rows of similarity matrices. For instance, the transition matrix from prototypes to points is obtained by softmaxing A^T, $\Gamma^{(\mathbf{p}\to x)} = \text{softmax}(A^T)$, such that $p(x_i|\mathbf{p}_j) = \Gamma^{(\mathbf{p}\to x)}_{j,i}$. Similarly, transition from points to prototypes $\Gamma^{(x\to\mathbf{p})}$, and transitions between points $\Gamma^{(x\to x)}$, are computed by softmaxing A, and B, respectively. Now, we define our random walker matrix as

$$T^{(\tau)} = \Gamma^{(\mathbf{p}\to x)} \cdot (\Gamma^{(x\to x)})^{\tau} \cdot \Gamma^{(x\to\mathbf{p})}, \quad (3)$$

where τ denotes the number of steps taken between the unlabeled points, before stepping back to a prototype. An entry $T_{i,j}$ denotes the probability of ending a walk at prototype j *given* that we have started at prototype i, and the j-th row is the probability distribution over ending prototypes, given that we started at prototype j. The diagonal entries of T denote the probabilities of returning to the starting prototype; our landing probabilities. Our goal is to maximize those by minimizing a cross-entropy loss between the identity matrix I and our random walker matrix T, dubbed as L_{walker}[1]

$$\begin{aligned}\mathcal{L}_{walker} &= \sum_{i=0}^{\tau} \alpha^i \cdot H(I, T^{(i)}), \quad \mathcal{L}_{visit} = H(\mathcal{U}, P),\\ \mathcal{L}_{RW} &= \mathcal{L}_{walker} + \mathcal{L}_{visit},\end{aligned} \quad (4)$$

where $H(I,T) = -\frac{1}{N_c}\sum_{i=0}^{N_c} \log T_{i,i}$, and α is an exponential decay hyperparameter. However, one issue with L_{walker} loss, is that we could end up visiting a small subset of the unlabeled points. To remedy this problem, Haeusser et al. (2017a) introduce a 'visit loss', pressuring the walker to visit a large set of unlabeled points. Hence, we assume that our walker is equally likely to start at any prototype, then we compute the overall probability that each point would be visited when we step from prototypes to points. $P = \frac{1}{N_c}\sum_{i=0}^{N_c} \Gamma_i^{(\mathbf{p}\to x)}$, where $\Gamma_i^{(\mathbf{p}\to x)}$ represents a column of the matrix. Then we add $\mathcal{L}_{visit}$ as the standard cross-entropy between this probability distribution and the uniform distribution $\mathcal{U}$. Hence, our final random walk loss is $\mathcal{L}_{RW}$ is the sum of $\mathcal{L}_{walker}$ and $\mathcal{L}_{visit}$; see Eq 4.

1. To be exact, this is the average cross-entropy between the individual rows of I and T, since those are probability distributions.

Overall Loss. To put it all together, our objective function could be written as $\arg\min_\theta \mathcal{L}_S + \lambda\mathcal{L}_{RW}$, where λ is a regularization parameter. While gradient of $\mathcal{L}_S = -\sum_{i=0}^{Q_L} y_i \log z_{i,c}$ provides the supervised signal, the gradient of $\mathcal{L}_{RW}$ encourages the *"prototypical magnetization"* property guided by our random walk. This loss is minimized in expectation over randomly sampled semi-supervised episodes from our training data.

3. Related Work

Associative learning and local consistency in Semi-Supervised Learning SSL contains a rich toolbox of principles and techniques to leverage unlabeled data to learn better discriminative embeddings. The core idea is that similar inputs tend to close in embedding space, measured with metrics such as Euclidean distance, and KL-divergence. Loss functions are designed to further encourage inputs belonging to the same class to cluster together in embedding space, such as triplet loss, mean teacher and learning by association (Haeusser et al., 2017b).

Particularly, our work can be seen as a generalization of learning by association, which is a special case of PRWN with a walk step size of 1. By using a walk step size larger than 1, we effectively increase the local receptive field to further associate and cluster similar inputs together in embedding space

Semi-Supervised Few-Shot Learning In a few-shot learning task, the model is assumed to be meta-learned (read: pretrained) to reach a good starting point, so that in the testing phase, learning with very few samples possible (Miyato et al., 2018; Kamnitsas et al., 2018; Haeusser et al., 2017a).

However, few-shot learning assumes the dataset to be meta-learned is sufficiently large and fully labeled. Hence, Ren et al. (2018) introduced the SS-FSL setting, so that FSL is feasible where labeled data is scarce. This setting combines both SS and FSL settings, where the meta-learning phase leverage unlabeled data to better meta-learn the task.

SS-FSL is also applied in other ways such as self-teaching (Li et al., 2019) and adaptive subspace (Simon et al., 2020), and applied in other tasks, such as transfer learning (Yu et al., 2020) and image translation (Wang et al., 2020). Orthogonal to these developments, our goal is to show that learning representations can be efficiently improved by prototypical random walk loss.

Application of Random Walk for Associative Learning Our work is also similar to the application of random walk for person re-identification (Shen et al., 2018), where a random walk is used to re-rank and find the best match given input *probe image* relative to the collection of known *gallery images*. However, our focus is on applying PRWN for semi-supervised few-shot learning tasks, that is to leverage unlabeled data for meta-training the model to good initial parameters, so that it is able to learn with few samples during test time.

4. Experiments on 2D synthetic datasets

To gain an intuition on how the proposed method works, we performed experiments on 2D synthetic datasets to easily visualize how the decision boundary is formed.

The model is 3-layer MLP, 2-dimension input, 32-dimension hidden unit, 4-dimension output. It used a negative Euclidean distance metric for its output. We used two datasets, and 3 models trained 300 epochs in each data- set: (1) Spiral dataset with 1000 points split to 7 labels (10% labels + Random Walk, 10% labels, and 100% labels; 1 shot, 5 way, $\tau = 1$); (2) Gaussian circle dataset 1000 points split into 3 labels. There were 3 models trained in each dataset (5% labels + Random Walk, 5% labels, and 5%

labels; 1 shot 3 way, $\tau = 1$) ; *cf.* Fig. 2 shows the results, it can be seen that the proposed method can "connect the dots" of unlabeled points in the green region and purple region, hence producing decision boundary similar to 100% labels. In the Gaussian circle dataset, random walk loss helps the model fits the circle more in just a few epochs, but the model without random walk loss still has many mis-classified points and the circular outline is not obvious.

5. Experiments on Image Datasets

Overview. In these experiments, we cover two main results: with and without distractors, where distractors are present at train *and* test time when applied. In each, we discuss experiments with and without *semi-supervised adaptation* where additional unlabeled data are used at *test* time. Note that whether or not unlabeled data is available at test time, we use the same trained model, the difference comes from adding the adaptation step in Eq. 2 at test time to leverage that data.

5.1. Experimental Setup

Datasets. We evaluated our work on the two commonly used SS-FSL benchmarks Omniglot, Mini-ImageNet, and tiered-ImageNet. Omniglot (Lake et al., 2011) is a dataset of 1,623 handwritten characters from 50 alphabets. Each character was drawn by 20 human subjects. We follow the few-shot setting proposed by Vinyals et al. (2016), in which the images are resized to 28×28 px and rotations in multiples of 90° are applied, yielding 6,492 classes in total. These are split into 4,112 training classes, 688 validation classes, and 1,692 testing classes. Mini-ImageNet (Vinyals et al., 2016) is a modified version of the ILSVRC-12 dataset (Russakovsky et al., 2015), in which 600 images, of size 84×84 px, for each of 100 classes were randomly chosen to be part of the dataset. We rely on the class split used by Ravi and Larochelle (2017a). These splits use 64 classes for training, 16 for validation, and 20 for testing. tiered-ImageNet (Ren et al., 2018) is also a subset of the ILSVRC-12 dataset (Russakovsky et al., 2015). However, it is way bigger than the Mini-ImageNet dataset in the number of images; around 700K images, and the number of classes; around 608 classes coming from 34 high-level categories. Each high-level category has about 10 to 20 classes and split into 20 training (351 classes), 6 validation (97 classes) and 8 test (160 classes) categories.

In our experiments, following Ren et al. (2018); Zhang et al. (2018), we sample 10% and 40% of the points in each class to form the labeled split for Omniglot and Mini - Imagenet, respectively; the rest forms the unlabeled split.

Implementation Details. We have provided full details of our experimental setting including network architectures, hyperparameter tuning on the validation set in supp. materials.For a fair comparison, we opt for the same Conv-4 architecture (Vinyals et al., 2016) appeared in the prior SS-FSL art (Zhang et al., 2018; Ren et al., 2018).

Episode Composition. All testing is performed on 5-way episodes for both datasets. Unless stated otherwise, the analysis performed in sections 5.2 & 5.3 are performed by averaging results over 300 5-shot 5-way mini-imagenet episodes from the *test* split, with N_u=10. Further detail is in supp. materials. All accuracies reported are averaged over 3000 5-way episodes and reported with 95% confidence intervals.

Baselines. We evaluate our approach on standard SS-FSL benchmarks and compare to prior art; PN (Ren et al., 2018), Meta-GAN (Zhang et al., 2018) and EGNN-Semi (Kim et al., 2019). We also compare PRWN with 3 control models; the vanilla prototypical net-

work (PN) trained on the fully labeled dataset, denoted PN_{all} (the oracle), which is considered to be our *target* model, a PN (Ren et al., 2018) model trained only on the labeled split of the data (40% of the labels), which is essentially PRWN without our random walk loss, and finally a PN trained with the state-of-the-art VAT (Miyato et al., 2018) and entropy minimization as a strong baseline; we denote it as PN_{VAT}.

5.2. Semi-supervised meta-learning without distractors

For experiments without semi-supervised adaptation, we observe from the third horizontal section of Table 1, that PRWN improves on the previous state-of-the-art MetaGAN (Zhang et al., 2018), and EGNN-Semi (Kim et al., 2019) on all experiments, with a significant improvement on 5-shot mini-imagenet. It is worth mentioning that our PRWN has less than half the trainable parameters of MetaGAN which employs an additional larger generator.

Experiments with semi-supervised adaptation are presented in the bottom section in Table 1. Note that PRWN already improves on prior art without the adaptation. With the added semi-supervised adaptation, PRWN improves significantly, and the gap widens. On the 5-shot mini-imagenet task, PRWN achieves a relative improvement of 8,17%, 4,86%, and 8,28% over the previous state-of-the-art, (Ren et al., 2018; Liu et al., 2019; Kim et al., 2019), respectively. Similar behavior has been observed on tiered-ImageNet dataset outperforming existing methods in 1-shot classification and similar performance on 5-shot classification; note that standard deviation for (Kim et al., 2019) is not reported for 1 and 5-shot classification.

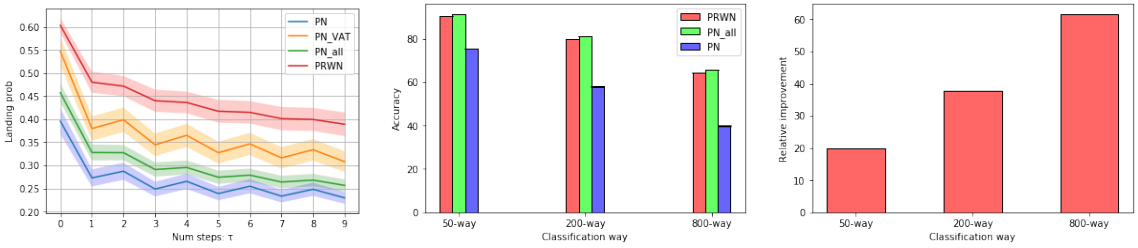
Ablation study. From Table 1, we can see that our PRW loss improves the baseline PN significantly, boosting the accuracy of PRWN up to 67.82% from 59.08% on the 5-shot mini-imagenet for example. Moreover, while PN_{VAT} proves a powerful model, competing with prior state-of-the-art, PRWN still beats it on all tests. Furthermore, We trained PRWN on mini-imagenet with only 20% of the labels, and we obtain an accuracy of 64.8% on the 5-shot task; outperforming the SOTA of 64.43% which uses double the number of labels. Most remarkably, PRWN performs competitively with the fully labeled PN_{all}, even outperforming it on 1-shot mini-imagenet.

Local & Global consistency Analysis. To evaluate the global consistency, we take a look into the behavior of our random walker for our various models. We compute the landing probability over the graphs they generate: the probability a random walker returns to the starting prototype, given by $Trace(T^{(\tau)})$ from Eq. 3. We can see in Fig. 2 that even as τ grows, PRWN generates graphs with the highest landing probs. Following is PN_{VAT}, implying that enforcing local consistency still helps with global consistency. We can also see that PN_{all} also does better than PN, indicating that the addition of extra labeled data also improves global consistency. To evaluate the local consistency and adversarial robustness of our various models, we compute their average VAT loss. Unsurprisingly, PN_{VAT} performs best with 1.1 loss, following are both PRWN and PN_{all} with 3.1 & 2.91 respectively, then PN with 5.9. We see again that improving global consistency helps with local consistency, and so does additional labeled data.

Discriminative Power. In order to study our approach and baselines in a more challenging setup, we evaluate their performance on a Higher-Way classification. Fig. 2 shows that our model still performs better than the baseline and close to PN_{all} (the oracle). The accuracy of PRWN, PN_{all}, and PN, on 800-ways in Omniglot, are 64.43%, 65.57% and 39.84%, respectively. In Fig. 2,

Table 1: Semi-Supervised Meta-Learning + Ablation Study

Model	Omniglot	Mini-Imagenet		Tiered-Imagenet	
	1-shot	1-shot	5-shot	1-shot	5-shot
PN_{all}(Snell et al., 2017)	98.8	49.4	68.2	53.6	74.34
PN (Ren et al., 2018)	94.62 ± 0.09	43.61 ± 0.27	59.08 ± 0.22	46.52± 0.52	66.15± 0.22
MetaGAN (Zhang et al., 2018)	97.58 ± 0.07	50.35 ± 0.23	64.43 ± 0.27	N/A	N/A
EGNN-Semi (Kim et al., 2019)	N/A	N/A	62.52 ± N/A	N/A	**70.98** ± N/A
PN_{VAT} (Ours)	97.14 ± 0.16	49.18 ± 0.22	66.94 ± 0.20	N/A	N/A
PRWN (Ours)	**98.28 ± 0.15**	**50.89 ± 0.22**	**67.82 ± 0.19**	**54.87 ± 0.46**	70.52 ± 0.43
PN + Semi-supervised inference(Ren et al., 2018)	97.45 ± 0.05	49.98 ± 0.34	63.77 ± 0.20	50.74 ± 0.75	69.37 ± 0.26
PN + Soft K-means(Ren et al., 2018)	97.25 ± 0.10	50.09 ± 0.45	64.59 ± 0.28	51.52 ± 0.36	70.25 ± 0.31
PN + Soft K-means + cluster(Ren et al., 2018)	97.68 ± 0.07	49.03 ± 0.24	63.08 ± 0.18	51.85 ± 0.25	69.42 ± 0.17
PN + Masked soft K-means(Ren et al., 2018)	97.52 ± 0.07	50.41 ± 0.24	64.39 ± 0.24	52.39 ± 0.44	69.88 ± 0.20
TPN-Semi (Liu et al., 2018)	N/A	52.78 ± 0.27	66.42 ± 0.21	55.74 ± 0.29	71.01 ± 0.23
PRWN + Semi-supervised inference (Ours)	**99.23 ± 0.08**	**56.65 ± 0.24**	**69.65 ± 0.20**	**59.17 ± 0.41**	**71.06 ± 0.39**



(a) Landing probabilities (b) Higher-Way Performance (c) Improvement over PN

Figure 2: *(a)* Landing Probabilities on mini-ImageNet: The x-axis denotes the number of steps for the walk (τ), and the y-axis shows the probability of returning to the right prototype. *(b):* The Higher-Way performance on Omniglot as we increase the number of test classes N_c. *(c):* The relative improvement of PRWN over PN as we increase the number of classes in Omniglot

we show the relative improvement over PN reaching $\approx 60\%$ improvement on 800-ways classification. Similar behavior has been reported for mini-imagenet (See Supp. materials).This shows the performance gain from our PRW loss is robust and reflects its discriminative power.

Transductive/Semi-supervised adaptation approaches. Our approach is orthogonal and can be integrated with these methods (Liu et al., 2019; Ren et al., 2018; Douze et al., 2018). In fact, PRWN + semi-supervised inference is such an integration where K-means step is integrated from (Ren et al., 2018). Tables 1, and 2 show that our network, combined with the K-means step at test time, performs far better than the networks trained with those adaptation methods. This supports our hypothesis that semi-supervised adaptation like the K-means step fails to fully exploit the unlabeled data **during meta-training**.

5.3. Semi-supervised meta-learning with distractors

The introduction of distractors by Ren et al. (2018) was meant to make the whole setup more realistic and challenging. To recap, the distractors are unlabeled points added to your support set, but they do not belong to any of the classes in that set i.e. the classes you are

Table 2: Experiments with distractor classes

Model	Omniglot	Mini-Imagenet		Tiered-Imagenet	
	1-shot	1-shot	5-shot	1-shot	5-shot
PRWN (Ours)	97.76 ± 0.11	50.96 ± 0.23	67.64 ± 0.18	53.30 ± 1.02	69.88 ± 0.96
PN+ Semi-supervised inference (Ren et al., 2018)	95.08 ± 0.09	47.42 ± 0.33	62.62 ± 0.24	48.67 ± 0.60	67.46 ± 0.24
PN+ Soft K-means (Ren et al., 2018)	95.01 ± 0.09	48.70 ± 0.32	63.55 ± 0.28	49.88 ± 0.52	68.32 ± 0.22
PN+ Soft K-means + cluster (Ren et al., 2018)	97.17 ± 0.04	48.86 ± 0.32	61.27 ± 0.24	51.36 ± 0.31	67.56 ± 0.10
PN+ Masked soft K-means (Ren et al., 2018)	97.30 ± 0.30	49.04 ± 0.31	62.96 ± 0.14	51.38 ± 0.38	69.08 ± 0.25
TPN-Semi (Liu et al., 2018)	N/A	50.43 ± 0.84	64.95 ± 0.73	53.45 ± 0.93	**69.93 ± 0.80**
PRWN+ Semi-supervised inference (Ours)	**97.86 ± 0.22**	**53.61 ± 0.22**	**67.45 ± 0.21**	**56.59 ± 1.13**	69.58 ± 1.00
PRWN+ Semi-supervised inference + filter (Ours)	**99.04 ± 0.18**	**54.51 ± 0.23**	**68.77 ± 0.20**	**57.97 ± 1.12**	69.74 ± 1.10

currently classifying over. This "labeled/ unlabeled class mismatch" was found by Oliver et al. (2018) to be quite a challenge for SSL methods, sometimes even making the use of unlabeled data harmful for the model. We present our results in table 2, where the top row is our model without test time adaptation, and we can see that it already beats the previous state-of-the-art below, which makes use of test time unlabeled data, even by a large margin in the 5-shot mini-imagenet with a relative improvement of 3,8%, and 6,1% on TPN-Semi (Liu et al., 2019), and PN+Soft K-Means (Ren et al., 2018), respectively. Moreover, it beats the MetaGAN (Zhang et al., 2018) model trained without distractors on all tasks, and in fact performs closely to our own PRWN trained without distractors (*cf.* Table 1).

When we add the semi-supervised adaptation step, with distractors present among unlabeled data at test time, we see that our model does not benefit well from that step, and in the case of the 5-shot mini-imagenet, the performance is slightly harmed. Next subsection, we will explore why our model is robust to distractors during training, and how we can use the random walk dynamics to make the semi-supervised inference step useful when distractors are present.

Distractor analysis

We hypothesize that the reason our PRWN is robust against distractors, is because our random walker learns to largely avoid distractor points, and as such they are not magnetized towards our class prototypes; if anything by learning to avoid them, the network is structuring the latent space such that points of each class are compact and well separated. This comes as a by-product of the "prototypical magnetization" property that our loss models.

To test this hypothesis, we take a PRWN model trained with distractors, we sample test episodes including distractors ($N_d = N_c = 5$), construct our similarity graph, and compute the probability that our random walker visits distractor versus non-distractor points. Concretely, we compute $P = \frac{1}{N_c}\sum_{i=0}^{N_c}\Gamma^{(\mathbf{p}\to x)}$, where the summation is over the columns, and an entry P_i represents the probability of visiting point i. We split P into P_{clean} and P_{dist}, containing the entries for non-distractor and distractor points. respectively. Both probabilities p_{clean} and p_{dist} should sum up to one. Whereas our baseline PN gets $p_{clean} = 0.67$, and PN_{all} gets $p_{clean} = 0.76$, our PRWN gets $p_{clean} = 0.81$. So we see our $\mathcal{L}_{RW}$ is not only an attractive force bringing points closer to prototypes, but it also has a repelling force driving irrelevant points away from prototypes. Note this is not only a feature of the network, it is a property of the loss function. For instance, the semi-supervised inference step (Ren et al., 2018) involves all points, distractor or not, equally in the prototype update, regardless of the geometry of the embeddings.

Distractors at semi-supervised inference. We also performed an experiment to further improve PRWN with the semi- supervised inference step. We exploit our random walk dynamics to order to filter out distractors. We compute the probability that a point is part of a successful walk; a walk which starts and ends at the same prototype. This is given by $S = \sum_{i=0}^{N_c} \Gamma^{(\mathbf{p} \to x)} \odot \Gamma^{(x \to \mathbf{p})}$, where $\odot$ is the Hadamard product, and the summation is over the columns of the resulting matrix. Then we simply discard the points that scored below the median. With this little step, we see our PRWN + semi-supervised inference, become more robust to test time distractors, with 99.04% accuracy on omniglot, 54.51% & 68.77% on mini - imagenet, and 57.97% & 69.74% on tiered - imagenet 1&5-shot, respectively. This simple filtering step just improved on the distractor state-of-the-art as shown in Table 2 (last row). Note that our approach also outperform Liu et al. (2019) by a significant margin in 1-shot classification in all datasets and 5-shot classification Mini-Imagenet, while achieving similar performance on 5-shot Tiered- Imagenet.

6. Conclusion

SS-FSL is a relatively unexplored yet challenging and important task. In this paper, we introduced a state-of-the-art SS - FSL model, by introducing a semi-supervised meta-training loss, namely the Prototypical Random Walk, which enforces global consistency over the data manifold, and magnetizes points around their class prototypes. We show that our model outperforms prior art and rivals its fully labeled counterpart in a wide range of experiments and analyses. We contrast the effects and performance of global versus local consistency, by training a PN with VAT (Miyato et al., 2018) and comparing it with our model. While the local consistency loss has an improvement on the performance, we found out that our global consistency loss significantly improves the performance in SS-FSL. Finally, we show that our model is robust to distractor classes even when they constitute the majority of unlabeled data. We show how this is related to the dynamic of PRW. We even create a simple distractor filter, and show its efficiency in improving semi-supervised inference (Ren et al., 2018). Our experiments and results set the state-of-the-art on most benchmarks.

References

Matthijs Douze, Arthur Szlam, Bharath Hariharan, and Hervé Jégou. Low-shot learning with large-scale diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3349–3358, 2018.

P. Haeusser, A. Mordvintsev, and D. Cremers. Learning by association — a versatile semi-supervised training method for neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 626–635, July 2017a. doi: 10.1109/CVPR.2017.74.

Philip Haeusser, Alexander Mordvintsev, and Daniel Cremers. Learning by association–a versatile semi-supervised training method for neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–98, 2017b.

Konstantinos Kamnitsas, Daniel C. Castro, Loïc Le Folgoc, Ian Walker, Ryutaro Tanno, Daniel Rueckert, Ben Glocker, Antonio Criminisi, and Aditya V. Nori. Semi-supervised learning via compact latent space clustering. In *ICML*, 2018.

Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. Edge-labeling

graph neural network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2019.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. In *Advances in Neural Information Processing Systems*, pages 10276–10286, 2019.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv preprint arXiv:1805.10002*, 2018.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *ICLR*, 2019.

Erik G Miller, Nicholas E Matsakis, and Paul A Viola. Learning from one example through shared densities on transforms. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 464–471. IEEE, 2000.

Takeru Miyato, Shin-ichi Maeda, Shin Ishii, and Masanori Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

Avital Oliver, Augustus Odena, Colin Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of semi-supervised learning algorithms. 2018. URL https://arxiv.org/pdf/1804.09170.pdf.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations*, 2017a.

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

Mengye Ren, Sachin Ravi, Eleni Triantafillou, Jake Snell, Kevin Swersky, Josh B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HJcSzz-CZ.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2265–2274, 2018.

Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. Adaptive

subspaces for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4136–4145, 2020.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

Yaxing Wang, Salman Khan, Abel Gonzalez-Garcia, Joost van de Weijer, and Fahad Shahbaz Khan. Semi-supervised learning for few-shot image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4453–4462, 2020.

Zhongjie Yu, Lin Chen, Zhongwei Cheng, and Jiebo Luo. Transmatch: A transfer-learning scheme for semi-supervised few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12856–12864, 2020.

Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In *Advances in Neural Information Processing Systems*, pages 2371–2380, 2018.