Reconstruction guided Meta-learning for Few Shot Open Set Recognition

Sayak Nag, Dripta S. Raychaudhuri, Sujoy Paul, and Amit K. Roy-Chowdhury, *Fellow, IEEE*

Abstract—In many applications, we are constrained to learn classifiers from very limited data (few-shot classification). The task becomes even more challenging if it is also required to identify samples from unknown categories (open-set classification). Learning a good abstraction for a class with very few samples is extremely difficult, especially under open-set settings. As a result, open-set recognition has received limited attention in the few-shot setting. However, it is a critical task in many applications like environmental monitoring, where the number of labeled examples for each class is limited. Existing few-shot open-set recognition (FSOSR) methods rely on thresholding schemes, with some considering uniform probability for open-class samples. However, this approach is often inaccurate, especially for fine-grained categorization, and makes them highly sensitive to the choice of a threshold. To address these concerns, we propose Reconstructing Exemplar-based Few-shot Open-set ClaSsifier (ReFOCS). By using a novel exemplar reconstruction-based meta-learning strategy ReFOCS streamlines FSOSR eliminating the need for a carefully tuned threshold by learning to be self-aware of the openness of a sample. The exemplars, act as class representatives and can be either provided in the training dataset or estimated in the feature domain. By testing on a wide variety of datasets, we show ReFOCS to outperform multiple state-of-the-art methods.

Index Terms—Few-Shot Learning, Open-Set Recognition, Out-of-distribution detection, Meta-learning

1 Introduction

Deep neural networks have achieved excellent performance on a wide variety of visual tasks [1], [2], [3]. However, the majority of this success has been realized under the closed-set scenario, where it is assumed that all classes that appear during inference are present in the training set. In real-world applications, it is often difficult to obtain samples that exhaustively cover all possible semantic categories [4]. This inherent open-ended nature of the visual world restricts the wide-scale applicability of deep models and machine learning models in general. Thus, it is more realistic to consider an *open-set* scenario [5], where the predictive model is expected to not only recognize samples from the seen classes, but also to recognize when it encounters an *out-of-distribution* sample and reject it, rather than making a prediction for it.

Notable approaches for open-set recognition involve adversarial training [6] to reject adversarial samples which are too hard to classify, inducing self-awareness in CNNs [7] to reject out-of-distribution samples, and using extreme value statistics to re-calibrate classification scores of samples from novel classes [8]. All of these approaches require large amounts of labeled data per category for the seen classes. On the contrary, humans can easily grasp new concepts with very limited supervision and simultaneously perceive the occurrence of unforeseen abnormalities. Aiming to emulate this, we seek to perform open-set recognition in the *few-shot learning* scenario, which has been largely ignored in the literature. A visual description of the problem setting is shown in Fig. 1.

The challenge in few-shot open-set recognition (FSOSR) stems from the limited availability of samples for in-distribution classes. This complicates learning a good abstraction of the provided categories to comprehensively distinguish between low-likelihood in-distribution samples and actual out-of-distribution samples. Following the success

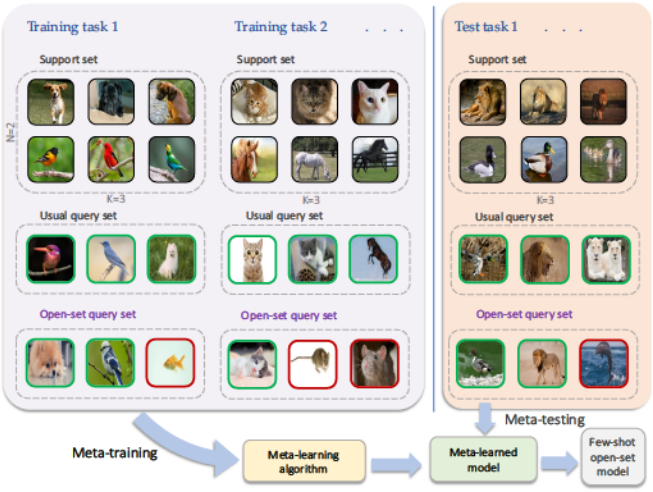

Fig. 1: **Problem setup.** We formulate few-shot open-set recognition as a meta-learning problem where training proceeds in an episodic manner. In contrast to the usual setup, where the support and query sets share the same categories, we consider the more challenging open-set scenario where the query set can contain samples from classes not seen in the support (highlighted in red).

of meta-learning approaches [9], [10], [11] in closed-set few-shot recognition, recent works [12], [13] attempt FSOSR by building on top of the popular Prototypical Network framework [11]. In [12], the authors employ entropy maximization to enforce uniform predictive distribution of out-of-distribution samples to maximize model confusion for unseen categories. Out-of-distribution samples are identified by thresholding the maximum probability of the classification logits. In contrast, [13] leverages a transformer [14] and

Bird	0.96	0.02	0.11	0.14
Dog	0.02	0.97	0.03	0.82
Truck	0.02	0.01	0.86	0.04

In-distribution Out-of-distribution

Fig. 2: **Assuming uniform predictive distribution for outliers.** Out-of-distribution samples can be correlated to in-distribution classes by varying amounts. In this figure, we train a classifier for three in-distribution classes: dog, bird and truck (corresponding samples highlighted in green). The out-of-distribution sample - a cat - shares highly similar visual characteristics to a dog, as evidenced by the prediction. This suggests that desiring outliers to have a uniform predictive distribution, as suggested in [12], is often inaccurate.

Fig. 3: **Canonical Exemplars of real images.** Examples of real-world images of symbolic data such as traffic signs and brand logos (left column) along with their corresponding class-specific exemplar images (right column). These exemplars are readily available for such symbols and lie on a canonical domain devoid of perturbations of the real images.

carefully crafted normalization techniques [13] to enforce improved representation consistency of in-distribution samples, enabling the rejection of out-of-distribution samples via simple distance thresholding in the feature space. Despite their effectiveness, these approaches rely heavily on carefully tuned threshold values. Furthermore, assuming near uniform predictive distribution for out-of-distribution samples does not generally hold in practice due to the sensitivity of deep CNNs to minor perturbations in the input [15] along with the tendency of the softmax operator to produce high confidence false predictions for out-of-distribution samples [7]. This is further explained in Fig. 2.

In order to overcome these drawbacks, we propose a different approach to detecting out-of-distribution samples in the few-shot setting, which utilizes *reconstruction* as an auxiliary task to induce self-awareness in a few-shot classifier. This self-awareness would allow the classifier to better detect when it is presented with an out-of-distribution sample, as it would be able to recognize when it is unable to accurately reconstruct the input. However, naively applying reconstruction fails in the few-shot setting due to overfitting. Inspired by [16], we propose using reconstruction of *class-specific exemplars*, instead of self-reconstruction, to flag out-of-distribution samples. Such exemplars act as ideograms to effectively encode the semantic information of the class it belongs to. Consequently, they serve to anchor the representations of in-distribution classes when access to a large number of samples is restricted. Many real-world graphical symbols, such as traffic signs and brand logos, have well-defined exemplars that lie on a simpler or canonical domain. Some examples are shown in Fig. 3. However, exemplars may not be available for all datasets, and for such cases, we provide a simple scheme to estimate them from the few-shot data in the *embedding space*, without any changes to the algorithmic formulation.

Building on this idea of reconstruction, we propose a new meta-learning strategy to tackle the few-shot open-set recognition task. In the meta-training phase, we use episodes sampled from a base set, with each episode simulating a token few-shot open-set task. Specifically, each episode is created by randomly selecting a set of classes and populating a support set with limited samples belonging to those classes. A query set is created in a similar fashion but contains samples from classes both seen in the support set and beyond (see Fig. 1). These episodes are subsequently used to train our framework: Reconstructing Exemplar-based Few-shot Open-set ClaSsifer (**ReFOCS**). Given an episode, ReFOCS projects these samples to a low-dimensional embedding space to perform a metric-based classification over the support classes. Simultaneously, a variational model is used to reconstruct the class exemplars to compute the reconstruction error. The class scores, in addition to the reconstruction error, are then utilized to recognize the probability of the sample being out-of-distribution.

Main Contributions

Our primary contributions are summarized below:

- We develop a new reconstruction based meta-learning framework which utilizes class-specific exemplars to jointly perform few-shot classification and out-of-distribution detection.
- By utilizing reconstruction as an auxiliary task in our meta-learning setup we induce self-awareness in our learning framework, enabling it to self-reject out-of-distribution samples without relying on carefully tuned thresholds.
- We introduce a novel embedding modulation scheme to make the learned representations more robust and discriminative in the presence of scarce samples. A weighted strategy for prototype computation is also introduced for reducing intra-class bias.
- Our framework outperforms or achieves comparable performance to the current state-of-the-art on a wide array of FSOSR experiments, thereby establishing it as a new baseline for FSOSR.

2 Related Works

Few-shot learning. Few-shot learning [17], [18], [19] aims to learn representations that generalize well to novel classes with few examples. Meta-learning [20] is the one of the most

common approaches for addressing this problem and it is generally grouped into one of the following categories: *gradient-based methods* [9], [21] and *metric learning methods* [10], [11]. Typical gradient based methods, such as MAML [9] and Reptile [21], aim to learn a good representation that enables fast adaptation to a new task. On the other hand, metric-based techniques like Matching Networks [10], and Prototypical Networks [11] learn a task-specific kernel function to perform classification via a weighted nearest neighbor scheme. Our framework belongs to the latter class of methods, with the ability to work in the open-set scenario.

Out-of-distribution detection. Also known as novelty or anomaly detection, the out-of-distribution task is commonly formulated as the detection of test samples that fall outside of the data distribution used in training. Hendrycks et. al. [7] showed that softmax alone is not a good indicator of out-of-distribution probability but statistics drawn from softmax can be utilized to make assumptions of the "normalcy" of a test sample. Liang et. al. [22] re-calibrated output probabilities by applying temperature scaling and used virtual adversarial perturbations to the input to enhance the out-of-distribution capability of the model. Note that most out-of-distribution detection focuses on either detecting perturbed samples or out-of-dataset samples. In addition, these methodologies are not extendable to few shot settings.

Open-set classification. Open-set classification is slightly different from out-of-distribution detection in that it focuses on not only rejecting samples from unseen classes but also achieving proper classification of the seen categories. This is a much harder problem as opposed to just detecting perturbed/corrupted samples [12]. Bendale et al. [8] introduced OpenMax which uses extreme value statistics to re-calibrate the softmax scores of samples from unseen classes and reject out-of-distribution samples by thresholding their corresponding confidence score. G-OpenMax [23] combines OpenMax with a generative model to synthesize the distribution of all unseen classes. Recently, counterfactual image generation [24] has been proposed to generate hard samples in an effort to build a more robust model. Kong et al. [25] extends this idea by designing a generative adversarial network for the generation of out-of-distribution samples. Wang et al. [26] followed a similar strategy to OpenMax [8] and introduced an energy-based classification loss for re-calibration of the softmax classification scores. Note that all these methods are in a fully supervised learning setting and are not directly applicable to few-shot learning. Liu et. al. [12] first address open-set recognition under the few-shot setting. The proposed framework, titled PEELER, builds on top of prototypical networks [11] and uses episodic learning guided entropy maximization for regularizing the softmax classification scores. Unseen categories are detected by thresholding the maximum value of the softmax score of each out-of-distribution sample. A more recent work SNATCHER [13] builds on top of PEELER [12] and utilizes transformers [14] along with carefully crafted normalization schemes to improve feature level consistency of in-distribution samples. This enables SNATCHER to detect samples from open-set categories by simple distance thresholding in the embedding space. These methods are, however heavily dependent on the choice of the threshold. To eliminate this dependency, we propose to address few-shot open-set recognition by utilizing exemplar reconstruction as an auxiliary task in the meta-learning setup, such that model is imbibed with self-awareness regarding the openness of an input sample. Exemplars have been used in many computer vision tasks ranging from image recognition [16] to panoptic segmentation [27]. Our work is the first to study there effectiveness on the FSOSR problem.

3 METHODOLOGY

In this section, we present our framework for open-set few-shot classification. We first provide a definition of the problem, followed by a brief overall methodology, and then present a detailed description of our framework.

Problem setting

Consider the standard few-shot learning setting, where we have access to a support set of labeled examples $\mathbb{S} = \{\mathrm{S}_1,...,\mathrm{S}_N\}$. Each $\mathrm{S}_c = \{\mathbf{x}_i\}_{i=1}^{K}$ denotes a set of K examples belonging to the class y_c, for N-way K-shot recognition. We make two changes to this setup. First, we assume the existence of class-specific exemplar images $\mathbf{t}_c$ for each $\mathbb{S}_c$; in case exemplars are not present, we estimate a class-specific exemplar for all the categories. Second, the query set $\mathbb{Q}$ is comprised of both in-distribution and out-of-distribution samples w.r.t $\mathbb{S}$, i.e., $\mathbb{Q} = \mathbb{Q}_{in} \cup \mathbb{Q}_{out}$. In-distribution samples belong to classes seen in the support set while out-of-distribution samples belong to unseen classes. The goal is to detect the samples in $\mathbb{Q}_{out}$ as out-of-distribution, while correctly classifying the samples in $\mathbb{Q}_{in}$.

In order to develop a strong prior for few-shot learning, we utilize a base training set of labeled samples $\mathcal{B} = \{(\mathcal{X}_c, \mathcal{Y}_c)\}_{c=1}^{M}$ to meta-train our framework, where M is the set of all classes in $\mathcal{B}$, $\mathcal{X}_c = \{\mathbf{x}_1, ..., \mathbf{x}_{|c|}\}$ is the set of images in class c and $\mathcal{Y}_c$ denotes the c^{th} class label. If class-specific exemplars $\mathbf{t}_c$ are provided, we add them to $\mathcal{B}$ to obtain $\hat{\mathcal{B}} = \{(\mathcal{X}_c, \mathcal{Y}_c, \mathbf{t}_c)\}_{c=1}^{M}$, otherwise we first estimate $\mathbf{t}_c$ from $\mathcal{B}$ as shown in section 3.5 and then add them to $\mathcal{B}$ to get $\hat{\mathcal{B}}$ in a similar fashion. Like prior work [10], [11] a set of training tasks or episodes $\{\mathcal{T}_1, \mathcal{T}_2, ...\}$ are sampled from $\hat{\mathcal{B}}$ to conduct meta-training of the model. For an N way K shot problem, each episode $\mathcal{T}$ is constructed by first randomly sampling N classes from B and then constructing a support set $\mathbb{S}$ comprised of K randomly chosen samples from each of the N classes. Along with $\mathbb{S}$, a query set $\mathbb{Q}$ is constructed in a similar fashion. In order to simulate the presence of out-of-distribution samples, we follow the strategy outlined in [12] and augment the query set with samples from classes *absent* in the support.

Overall framework

A pictorial description of our framework is shown in Fig. 4. Given a sample $\mathbf{x}$, we first use variational inference to reconstruct the possible exemplar $\mathbf{t}$ associated with the ground-truth class of $\mathbf{x}$ and simultaneously obtain a latent representation of the same sample. This embedding is used to obtain a classification score, similar to [10], [28], while the reconstructed exemplar is used as a proxy for the out-of-distribution detection task. Specifically, if $\mathbf{x} \in \mathbb{Q}_{out}$, we hypothesize that the reconstruction will fail

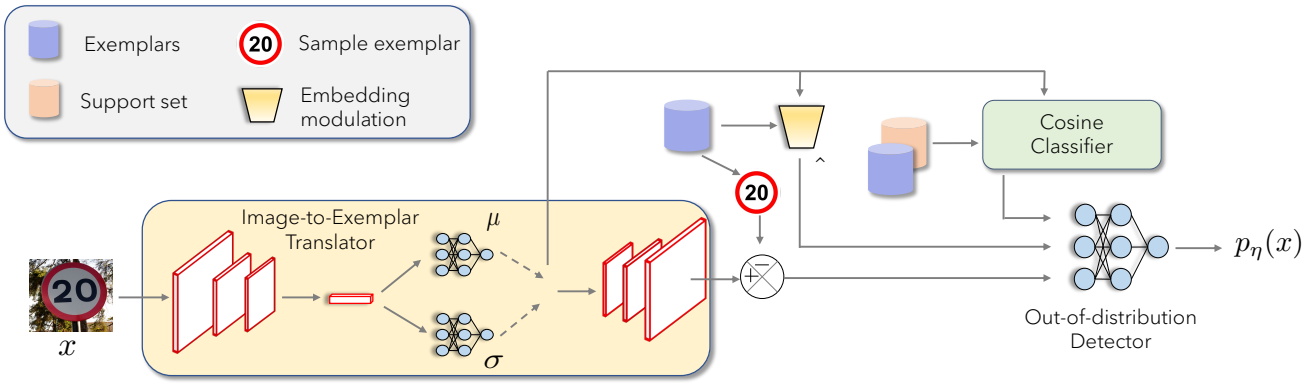

Fig. 4: **Overview of framework.** Given a query sample $\mathbf{x}$, a latent representation $\mathbf{z}$ is derived by sampling from the variational posterior, the parameters of which are μ and σ. This embedding is further enhanced via a modulation process to get $\hat{\mathbf{z}}$. The latent embedding is used for classifying the sample into one of the few-shot classes. The decoder reconstructs the exemplar, $\mathbf{t}$ associated with the sample's class. The exemplar reconstruction error, the modulated embedding, and the classification scores themselves are fed to an MLP-based out-of-distribution detector to predict the probability p_η of whether $\mathbf{x}$ is in/out-of-distribution with respect to the few-shot classes.

for all the exemplars of the support classes. Based on this hypothesis, the latent representation, classification scores, and reconstruction errors (with respect to support exemplars) are subsequently fed into a binary classifier to predict the probability of the sample $\mathbf{x}$ being out-of-distribution.

3.1 Exemplar reconstruction from real images

Since the class-specific exemplars $\mathbf{t}$ form a compact representation of the real world images belonging to that class, we hypothesize that the ability to reconstruct any exemplar belonging to in-distribution classes correlates positively with the sample being an in-distribution one. Inspired by [16], we use a Variational Auto-Encoder (VAE) [29] for the purpose of such reconstruction. The choice of VAE is motivated by its robustness to outliers and better generalization to unseen data, as shown in [30]. This is ideal for the image-to-exemplar translation task where the exemplars lie on a canonical domain devoid of the perturbations in real images.

Given an input sample $\mathbf{x}$, its exemplar reconstruction of $\mathbf{t}$ is carried out by maximizing the variational lower bound of the likelihood $p(\mathbf{t})$ [16] as follows,

$$\log p(\mathbf{t}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{t}|\mathbf{z})] - D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \quad (1)$$

where $D_{KL}[.]$ is the Kullback-Leibler (KL) divergence and $q_\phi(\mathbf{z}|\mathbf{x})$ denotes the variational distribution introduced to approximate the intractable posterior. Note that this is different from a vanilla VAE [29], which is derived by maximizing the log-likelihood of the input data $\mathbf{x}$. A differentiable version of the lower bound is derived by assuming the latent variable $\mathbf{z}$ to be Gaussian in nature, sampled from the prior $q_\phi(\mathbf{z}|\mathbf{x})$. The empirical loss to be minimized is given as follows,

$$\mathcal{L}_{VAE} = \frac{1}{K_e}\sum_{i=1}^{K_e} -\log p_\theta(\mathbf{t}_i|\mathbf{z}_i) + D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \quad (2)$$

where K_e is the number of samples in one episode, i.e., $K_e = |\mathbb{S} \cup \mathbb{Q}_{in}|$. Since $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ is non-differentiable, the re-parameterization trick is applied via the decoder network [29] such that $\mathbf{z} = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\odot$ is the Hadamard product.

The first term in Eq. 2 is the reconstruction loss which affects the mapping of the real images to their class-specific exemplars, while the second term acts as a distribution regularization, enforcing the latent variable $\mathbf{z}$ to follow the chosen prior. Although binary cross entropy (BCE) is the common choice for the reconstruction loss in VAE, other losses such as ℓ_1 or ℓ_2 norm can also be used.

3.2 Few-shot classification

The latent representation of a query sample $\mathbf{z}_q$, obtained from the encoder in the previous section, is used for computing the classification scores. We use the cosine metric to compute a relation score between the query sample and the support set $\mathbb{S}$. Specifically, the relation score is obtained by computing the cosine similarity between $\mathbf{z}_q$ and the set of prototypes or centroids, $\{\mathbf{\Omega}_c\}_{c=1}^{N}$ for each class $c \in \mathbb{S}$ (Fig. 4). The classification score for $\mathbf{x}_q$ is obtained by performing a softmax operation on these relation scores. In contrast to prior works [11], [12], the class specific prototype $\mathbf{\Omega}_c$ is obtained by a weighted mean of the support samples instead of a simple mean. Note that these prototypes are different from the class specific exemplars $\mathbf{t}$.

Prototype computation

Given the support set, the prototypes for each class c are calculated as follows,

$$\mathbf{\Omega}_c = \sum_{k=1}^{K} \omega_k \cdot \mathbf{z}_k^c \quad (3)$$

$\mathbf{z}_k^c$ denotes the latent representation of $\mathbf{x}^k \in \mathbb{S}_c$, while ω_k is the weight assigned to $\mathbf{z}_k^c$ based on how close it is from the embedding of the exemplar belonging to the c^{th} class, $\mathbf{z}_t^c$,

$$\omega_k = \frac{e^{\cos(\mathbf{z}_k^c, \mathbf{z}_t^c)}}{\sum_{k=1}^{K} e^{\cos(\mathbf{z}_k^c, \mathbf{z}_t^c)}} \quad (4)$$

We use these weights in an effort to control the phenomena of *intra-class bias* [31], i.e., the difference between the true expected prototype and the Monte-Carlo estimated

value. As explained earlier, each of the exemplars are an effective ideogram which provide a good abstraction of their respective classes. Thus the image-to-exemplar translation learned by the VAE leads to a feature space where the embedding of the real images cluster around that of their corresponding exemplars. This makes the exemplars good approximation of the true prototype and we leverage this via the weights ω_k to alleviate the intra-class bias.

Classification

After computing the prototypes, we predict the classification scores for the query sample x_q as follows,

$$p_\phi(y=c|\mathbf{x}_q)=\frac{e^{\tau\cdot cos(\mathbf{z}_q,\mathbf{\Omega}_c)}}{\sum\limits_{c'}e^{\tau\cdot cos(\mathbf{z}_q,\mathbf{\Omega}_{c'})}} \tag{5}$$

where τ is a learnable temperature parameter to scale the logits computed by cosine similarity [28], [32]. Learning proceeds by minimizing a cross-entropy loss over the in-distribution classes as follows:

$$\mathcal{L}_{CE}=-\frac{1}{|\mathbb{Q}_{in}|}\sum_{i=1}^{|\mathbb{Q}_{in}|}\sum_{c=1}^{N}\mathbb{1}\{y_i=c\}\log\, p_\phi(y=c|\mathbf{x}_{q,i}) \tag{6}$$

where y_i represents the true class of the query sample.

3.3 Out of distribution detection

Unlike prior work, we do not rely solely on the classification score to detect out-of-distribution query samples. Instead, ReFOCS flags query samples by leveraging the output of a multi-layer perceptron (MLP) classifier. This binary classifier takes into account three sources of information for scoring the openness of a query sample. These sources are (i) the class probability $\mathbf{p}_\phi$ as predicted in Eq. 5, (ii) a modulated version of the latent representation $\mathbf{z}$ (described below), and (iii) the set of reconstruction errors with respect to the support set exemplars, $\mathbf{D}=\left[||\hat{\mathbf{t}}-\mathbf{t}_1||_F^2,...,||\hat{\mathbf{t}}-\mathbf{t}_N||_F^2\right]$, where $\mathbf{D}$ indicates how far the reconstructed exemplar $\hat{\mathbf{t}}$ deviates from the actual exemplar $\mathbf{t}\in\mathbb{S}$. Intuitively, for out-of-distribution queries, all the entries of $\mathbf{D}$ will be very high, while for in-distribution samples, at least one of them will be very small.

Embedding Modulation

At a fine-grained level, samples from many of the out-of-distribution classes can have very similar visual features to some of the in-distribution classes (Fig 2. This issue becomes more relevant for the few-shot setting since the model does not have access to large amounts of samples from the in-distribution classes for generalization. Therefore, given only a handful of samples from the in-distribution classes, it is of paramount importance to obtain an embedding that is discriminative enough to provide good segregation between in-distribution and out-of-distribution classes. This, in turn, will help the MLP in better detecting the out-of-distribution samples. While the latent embedding obtained from the VAE does have good discriminative properties, we introduce an additional modulation step that can enhance it even further. The enhanced embedding, $\hat{\mathbf{z}}_\mathbf{q}$, is obtained by scaling the embedding of a query sample $\mathbf{z}_q$ with a scalar $\kappa>0$ as shown below,

$$\hat{\mathbf{z}}_\mathbf{q}=\frac{\mathbf{z}_q}{\kappa},\ \text{where}\ \kappa=\min_{c\,\in\,\mathbb{S}}||\mathbf{z}_q-\mathbf{\Omega}_c||_1 \tag{7}$$

where $c\in\{1,...N\}$ represents the classes in the support and $\mathbf{\Omega}_c$ is the prototype corresponding to the c^{th} class. κ is a modulation factor measuring how close a query sample is to any of the in-distribution classes in the embedding space [33]. out-of-distribution samples will tend to have higher values for κ compared to in-distribution samples. Therefore this form of modulation will amplify the embeddings of in-distribution samples while scaling them down for out-of-distribution queries.

Therefore, the final input to the out-of-distribution detecting MLP is the concatenated vector $[\mathbf{p}_\phi,\hat{\mathbf{z}}_q,\mathbf{D}]$. This MLP classifier outputs a sigmoidal probability value p_η, indicating the openness of a query sample. Training proceeds by minimizing binary cross-entropy loss as follows,

$$\mathcal{L}_{BCE}=-\frac{1}{|\mathbb{Q}|}\sum_{i=1}^{|\mathbb{Q}|}y_{\eta,i}\log\, p_{\eta,i}+(1-y_{\eta,i})\log(1-p_{\eta,i}) \tag{8}$$

where y_η is equal to 0 or 1 depending on whether $x_q\in\mathbb{Q}_{in}$ or $x_q\in\mathbb{Q}_{out}$ respectively.

3.4 Training

The parameters of the Encoder (ϕ), Decoder (θ) and the out-of-distribution detector (η) are jointly meta-trained by optimizing over the aggregate loss $\mathcal{L}$,

$$\mathcal{L}=\lambda_1\mathcal{L}_{VAE}+\lambda_2\mathcal{L}_{CE}+\lambda_3\mathcal{L}_{BCE} \tag{9}$$

where λ_1, λ_2 and λ_3 are hyper-parameters choices of which is discussed in Section 4.

3.5 Estimation of exemplars

While it is easy to obtain well-defined exemplar images for images of graphical symbols, such class-specific exemplars may not always be provided for all kinds of natural images. We perform a simple exemplar estimation for categories that have missing exemplar images. First, we perform non-episodic training of the VAE encoder, f_ϕ on the entire base set $\mathcal{B}$. Second, the category-wise samples in $\mathcal{B}$ are passed through the pre-trained encoder to obtain corresponding feature representations, $f_\phi(\mathbf{x}_c)$ from the penultimate layer of the encoder. Finally, the exemplar image is defined via a nearest neighbor scheme in the feature space as follows,

$$\mathbf{t}_c=\arg\min_{\mathbf{x}\in\mathcal{X}_c}||f_\phi(\mathbf{x})-\Psi_c||_2. \tag{10}$$

Here, $\Psi_c=\frac{1}{|\mathcal{X}_c|}\sum\limits_{\mathbf{x}\in\mathcal{X}}f_\phi(\mathbf{x})$ is the centroid of the c^{th} training class in the feature space.

For test episodes, we calculate the exemplar in a similar fashion by selecting the support sample closest to its centroid representation in the feature space.

4 Experiments

In this section, we provide comprehensive experiments over several data sets to prove the efficacy of our framework, ReFOCS, for few-shot open-set classification. We compare the performance of ReFOCS with existing state-of-the-art methods that rely on thresholding approaches and softmax re-calibration for the detection of out-of-distribution samples. The overall results show that our method outperforms or achieves comparable performance to existing methods, without relying on the need for carefully tuned thresholds.

TABLE 1: **Episodic sampling strategy for each of the few-shot experiments.** For each episode $\mathcal{T}$, $\mathbf{K}^c_{\mathbb{Q}_{in}}$ denotes the number of in-distribution queries sampled from each support class $c \in \{1,..,N\}$ and the total number of in-distribution query samples is $\mathbf{K}_{\mathbb{Q}_{in}} = \sum_{c=1}^{N} \mathbf{K}^c_{\mathbb{Q}_{in}}$. The total number of OOD samples for each episode is denoted as $\mathbf{K}_{\mathbb{Q}_{out}}$. Hence, the total number of samples in each episode is $\mathbf{K} + \mathbf{K}_{\mathbb{Q}_{in}} + \mathbf{K}_{\mathbb{Q}_{out}}$ where, $\mathbf{K} = |\mathbb{S}|$. The model is meta-trained with a total of $\mathbf{E}_{train}$ number of episodes and meta-tested on $\mathbf{E}_{test}$ number of episodes.

DATASETS	GTSRB→GTSRB		GTSRB→TT100K		Belga→Flickr32	Belga→Toplogos	*mini*ImageNet→*mini*ImageNet	
EXPERIMENT	5-way 5-shot	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 1-shot	5-way 1-shot	5-way 5-shot	5-way 1-shot
$\mathbf{K}^c_{\mathbb{Q}_{in}}$	10	10	10	10	1	1	15	15
$\mathbf{K}_{\mathbb{Q}_{out}}$	50	50	50	50	5	5	75	75
$\mathbf{E}_{train}$	20000	20000	20000	20000	50000	50000	50000	50000
$\mathbf{E}_{test}$	800	800	800	800	1700	400	600	600

Datasets

We use six datasets to set up three different types of few-shot open-set recognition experiments: (i) *traffic sign recognition*, for which we use the GTSRB [34] and TT100K [35] datasets, (ii) *brand logo recognition*, for which we use BelgaLogos [36], [37], FlickrLogos-32 [38] and TopLogo-10 [39], and (iii) *natural image classification* on the benchmark *mini*ImageNet dataset [10]. Different base training and meta-testing sets are configured from these datasets to obtain five few-shot scenarios as shown in Tables 2, 3 and 4. Some of these scenarios involve cross-dataset experimentation, which is more challenging compared to using splits from the same dataset and better mimics real-world scenarios where training and test data can have significant domain shifts [40]. It must be noted that all the traffic sign and logo datasets are provided with well-defined canonical exemplars for each of the classes. For these datasets the canonical exemplars can be directly used for the reconstruction task. To show the efficacy of our framework, we show results with both estimated and canonical exemplars for the traffic sign and logo datasets. The estimated exemplars are obtained following the strategy in section 3.5. Since *mini*ImageNet is not provided with any well-defined exemplars, all of its class-specific exemplars are estimated. For the traffic sign and natural image datasets, we evaluate our model on both the 5-way 5-shot and the 5-way 1-shot tasks, while for the logo classification task, we show results only on the 5-way 1-shot scenario. This is due to BelgaLogos and Toplogos having multiple classes with less than 5 samples. More details on these datasets and splits can be found in the supplementary material.

Implementation

The episodic sampling strategies for all the few-shot experiments are shown in Table 1. For the traffic sign and logo classification experiments, the VAE architecture is adapted from [16]. For the natural image classification task, we experiment with different resnet architectures [41] for the VAE encoder. In each case, the decoder is designed as an inverted version of the encoder architecture, for e.g., if Resnet18 [41] is used as the encoder, the decoder is designed as an inverted Resnet18. For all experiments, the MLP-based out-of-distribution detector is designed with two hidden layers, each containing 200 and 100 nodes, respectively. For a fair comparison, the encoder network is kept the same for all competing methods. The Adam optimizer [42] is used for all the experiments. For all the traffic sign and brand logo recognition experiments, the values of the loss function hyperparameters, λ_1,λ_2 and λ_3, are set to 10^{-4},10 and 10 respectively. For the natural image classification task on *mini*ImageNet λ_1, λ_2 and λ_3, are set to 10^{-4},1 and 10 respectively. The initial learning rate is set to 10^{-4} for both the traffic sign and brand logo experiments and to 10^{-3} for the natural image classification task. When using canonical exemplars for the traffic sign and brand logo experiments the standard BCE criterion is used as the VAE reconstruction loss. On the other hand, for all experiments involving the estimated exemplars, we observed using the ℓ_2 norm as the reconstruction loss results in more improved performance. Additional implementation details are provided in the supplementary.

Baselines

We compare ReFOCS against existing FSOSR methods such as PEELER [12] and SNATCHER [13]. Both these methods are built on top of the Prototypical Networks (ProtoNet) [11] and rely on thresholding to reject out-of-distribution samples. Therefore, we also compare the performance of ReFOCS to that of ProtoNet, which although not designed for open-set recognition, provides an approximate lower bound on FSOSR performance. We also compare against OpenMax [8], which was originally proposed for open-set recognition in the fully supervised setting. OpenMax fits a Weibull distribution on the classification logits and utilizes its parameters to recalibrate the final softmax classification score. In order to adapt it for the few-shot setting, we implement OpenMax over the standard ProtoNet. We henceforth denote this baseline as Proto+OM. For both ProtoNet and Proto+OM, the out-of-distribution samples are rejected by using a the same thresholding principle as PEELER.

Evaluation Metrics

To quantify the closed-set classification performance, we compute the accuracy over in-distribution queries and utilize the Area Under the Receiver Operating Characteristic curve (AUROC) to quantify the model's performance in detecting the out-of-distribution samples. For all our evaluations, we split the query set equally among in-distribution and out-of-distribution samples.

4.1 Traffic Sign Recognition

The performance of ReFOCS on the two traffic sign recognition tasks, GTSRB→GTSRB and GTSRB→TT100K, is shown in Table 2. GTSRB and TT100K have a total of 43 and 36 classes, respectively. For GTSRB→GTSRB, 22 classes are used for meta training, and the remaining 21 are used

TABLE 2: **5-way 5-shot and 5-way 1-shot results of traffic sign recognition**. For both traffic sign datasets GTSRB→GTSRB and GTSRB→TT100K, 800 test episodes were evaluated and the average performance is reported along with their 95% confidence levels.

Model	Metric	Exemplar	GTSRB→GTSRB		GTSRB→TT100K	
			Acc. (%)	AUROC(%)	Acc. (%)	AUROC (%)
			5-way 5-shot			
ProtoNet [11]	Euclidean	-	91.79 ± 0.44	70.74 ± 0.79	80.51 ± 0.78	62.10 ± 0.76
Proto+OM [8]	Euclidean	-	91.92 ± 0.43	86.67 ± 0.48	64.40 ± 0.78	68.80 ± 0.66
PEELER [12]	Mahalanobis	-	93.87 ± 0.37	90.99 ± 0.44	79.04 ± 0.79	73.25 ± 0.81
ReFOCS	Cosine	Estimated	93.89 ± 0.48	94.67 ± 0.25	80.09 ± 0.79	79.48 ± 0.62
ReFOCS	Cosine	Canonical	**94.17 ± 0.38**	**94.83 ± 0.35**	**83.36 ± 0.76**	**85.25 ± 0.56**
			5-way 1-shot			
ProtoNet [11]	Euclidean	-	82.31 ± 0.75	61.52 ± 0.90	69.82 ± 0.87	56.02 ± 0.83
Proto+OM [8]	Euclidean	-	82.46 ± 0.70	81.43 ± 0.65	55.10 ± 0.80	67.79 ± 0.66
PEELER [12]	Mahalanobis	-	82.86 ± 0.77	79.56 ± 0.85	**73.47 ± 0.88**	69.68 ± 0.88
ReFOCS	Cosine	Estimated	84.09 ± 0.71	**94.09 ± 0.21**	70.13 ± 0.89	75.01 ± 0.76
ReFOCS	Cosine	Canonical	**86.21 ± 0.78**	93.02 ± 0.45	71.45 ± 0.87	**81.98 ± 0.59**

TABLE 3: **5-way 1-shot results of brand logo recognition.** For **Belga→Flickr32**, 1700 test episodes were evaluated and for **Belga→Toplogos**, 400 test episodes were evaluated and their average closed-set Accuracy and open-set AUROC are reported with 95% confidence intervals.

Model	Metric	Exemplar	Belga→Flickr32		Belga→Toplogos	
			Acc. (%)	AUROC(%)	Acc. (%)	AUROC (%)
ProtoNet [11]	Euclidean	-	59.50 ± 0.99	58.69 ± 0.94	38.08 ± 2.10	53.18 ± 1.97
Proto+OM [8]	Euclidean	-	59.60 ± 1.01	61.00 ± 1.02	38.20 ± 1.95	56.40 ± 1.92
PEELER [12]	Mahalanobis	-	62.93 ± 1.03	66.40 ± 0.86	39.55 ± 2.05	56.25 ± 1.94
ReFOCS	Cosine	Estimated	65.15 ± 1.02	69.08 ± 0.84	41.35 ± 2.07	57.36 ± 1.94
ReFOCS	Cosine	Canonical	**66.29 ± 1.02**	**72.98 ± 0.83**	**42.30 ± 2.15**	**58.39 ± 1.97**

for meta testing. For GTSRB→TT100K, all classes of GTSRB are used for training, and testing is done on all the classes of TT100K. In both experiments, all images are resized to 64×64. As shown in Table 2, ReFOCS outperforms all baselines using both estimated and canonical exemplars. Specifically for the task of detecting out-of-distribution samples, on average, ReFOCS achieves **7.4** percentage points higher AUROC compared to PEELER while using the estimated exemplars. On the other hand, using the well-defined canonical exemplars yields even greater performance gains up to nearly **10.4** percentage points compared to PEELER. The problem of just thresholding softmax classification scores to indicate the openness of samples can be clearly seen in the lower AUROC values of ProtoNet, Proto+OM, and PEELER. The issue is more pronounced for ProtoNet, since, unlike PEELER and Proto+OM, ProtoNet does not explicitly regularize the probability scores of out-of-distribution samples resulting in even more high-confidence false predictions. Utilizing our proposed prototype computation (Eq. 4) also results in higher classification accuracy on most of the FSOSR tasks. Since the canonical exemplars lie in a simpler domain they are devoid of many background and lighting perturbations seen in the estimated exemplars. As a result, the image-to-exemplar translation is much simpler for canonical exemplars compared to the estimated ones, which results in higher reconstruction errors for the latter, in turn affecting the out-of-distribution detection. This factor is even more amplified if there is a domain shift between the base training and meta-testing sets as in the case of GTSRB→TT100K, which is why is comparison to GTSRB→GTSRB, using the estimated exemplars over the canonical ones leads to a larger drop in AUROC for GTSRB→TT100K. Nevertheless, the VAE-based reconstruction and our entire learning setup enable ReFOCs to better compensate for such domain shifts in comparison to existing FSOSR methods leading to significant increases in AUROC even while using the estimated exemplars. Some sample exemplar reconstructions for the canonical exemplar case are provided in Fig 5.

4.2 Brand Logo Recognition

The brand logo datasets consist of everyday images of commercial brand logos. In both few-shot tasks, Belga→Flickr32 and Belga→Toplogos, the Belga dataset consisting of 37 classes, is used for meta-training. Flickr32 and Toplogos each have 32 and 11 classes, respectively. Since some of the classes have as low as 2 samples, we restrict our experiments to the 5-way 1-shot scenario. In both cases, all images are resized to 64×64. Similar to the traffic sign recognition experiments, ReFOCS outperforms the existing baselines as evident from the results shown in Table 3. Specifically, while using

TABLE 4: **5-way 5-shot and 5-way 1-shot results of natural image classification**. The best results for each backbone are shown in bold. The performance of all methods is averaged over 600 test episodes and reported with 95% confidence levels. When using the Resnet12 backbone, it is pre-trained for all competing methods using the same regimen as [13]. † denotes our implementation using official code repository [1].

MODEL	BACKBONE	METRIC	*mini*IMAGENET→*mini*IMAGENET 5-way 5-shot		5-way 1-shot	
			ACC.(%)	AUROC(%)	ACC.(%)	AUROC(%)
PROTONET [11]	Resnet18	Euclidean	78.11 ± 1.22	58.67 ± 0.45	55.18 ± 1.36	56.51 ± 0.47
PROTO+OM [8]	Resnet18	Euclidean	79.01 ± 1.21	54.16 ± 0.47	55.73 ± 1.36	45.09 ± 0.39
PEELER [12]	Resnet18	Mahalanobis	75.08 ± 0.72	69.85 ± 0.70	**58.31 ± 0.58**	61.66 ± 0.62
PEELER† [12]	Resnet18	Mahalanobis	63.60 ± 0.67	64.17 ± 0.64	48.49 ± 0.81	59.39 ± 0.77
REFOCS	Resnet18	Cosine	**79.06 ± 1.17**	**69.91 ± 0.51**	**58.33 ± 0.84**	**69.24 ± 0.59**
PROTONET [11]	Resnet12	Euclidean	80.89 ± 1.18	59.14 ± 0.49	65.02 ± 1.33	52.47 ± 0.51
PROTO+OM [8]	Resnet12	Euclidean	80.33 ± 1.22	53.95 ± 0.51	65.14 ± 1.39	44.07 ± 0.36
PEELER† [12]	Resnet12	Mahalanobis	79.94 ± 0.63	73.58 ± 0.67	62.71 ± 0.79	64.87 ± 0.80
SNATCHER-F [13]	Resnet12	Euclidean	82.02	**77.42**	67.02 ± 0.85	68.27 ± 0.96
SNATCHER-T [13]	Resnet12	Euclidean	81.77	76.66	66.60± 0.80	70.17 ± 0.88
SNATCHER-L [13]	Resnet12	Euclidean	**82.36**	76.15	**67.60 ± 0.83**	69.40 ± 0.92
REFOCS	Resnet12	Cosine	**82.61 ± 1.14**	**77.31 ± 0.55**	66.29 ± 0.91	**71.23 ± 0.62**

TABLE 5: **Ablation studies.** Recons. w/ AE refers to swapping the VAE with an Autoencoder (AE); Non-weighted prototype refers to the computation of the prototype as a simple centroid similar to ProtoNet [11]; No Modulation denotes turning off embedding modulation; No embedding/clf denote the removal of embedding/softmax scores from the input of the novelty module; ProtoC+ND refers to when we donot use exemplars for anything and consequently remove the reconstruction errors from the input of the novelty module.

EXPERIMENT	GTSRB→TT100K 5-way 5-shot		BELGA→FLICKR32 5-way 1-shot	
	ACC.(%)	AUROC(%)	ACC.(%)	AUROC(%)
Recons. w/ AE	81.33 ± 0.80	77.08 ± 0.79	63.13 ± 1.05	71.86 ± 0.86
Non-weighted Prototype	81.79 ± 0.77	84.93 ± 0.59	64.07 ± 1.02	72.11 ± 0.86
No modulation	80.94 ± 0.79	78.55 ± 0.65	63.26 ± 1.03	56.69 ± 0.97
ProtoC+ND	76.92 ± 0.83	72.50 ± 0.61	61.64 ± 1.03	50.01 ± 0.99
No embedding	83.03 ± 0.75	74.33 ± 0.66	65.49 ± 1.00	67.79 ± 0.91
No clf	82.71 ± 0.74	82.85 ± 0.56	64.51 ± 1.01	71.58 ± 0.87
Full Model	**83.36 ± 0.76**	**85.25 ± 0.56**	**66.29 ± 1.02**	**72.98 ± 0.83**

estimated exemplars, on average, ReFOCS achieves nearly **1.89** percentage points higher AUROC compared to PEELER and with the canonical exemplars, the average AUROC gains are **4.28** percentage points. The overall accuracy is also significantly higher compared to the baselines. Specifically, when using the canonical exemplars we observe an average increase of **3.05** percentage points in accuracy over PEELER. These results establish the superiority of ReFOCS for FSOSR. Similar to the GTSRB→TT100K, the domain shift between BelgaLogos and the other two brand logo datasets increases the difficulty of the task, especially when using the estimated exemplars. However, the overall results clearly show the superiority of ReFOCS over existing FSOSR methods in compensating for such domain shifts.

4.3 Natural Image Classification

The *mini*ImageNet dataset, introduced in [10], is a benchmark dataset used for evaluating the performance of models for few-shot natural image classification. This dataset has a total of 100 classes, and as per the splits introduced in [10], 64 classes are used for training, 16 for validation, and the remaining 20 for testing. The images are resized to the standard resolution of 84×84 [10]. As mentioned previously, this dataset does not have any categorical exemplars, and therefore, for all its experiments, we use the estimated exemplars. As shown in Table 4, for this dataset, we show results with different encoder or backbone architectures. Specifically for the Resnet12 encoder, it is first pre-trained for all competing methods using the same setup as [13]. Following this, meta-training starts using the pre-trained weights of the encoder. From the overall results, we can observe that ReFOCS outperforms or achieves comparable performance to existing thresholding-based FSOSR methods. Since the images in *mini*ImageNet have much more variations, methods like Proto+OM fail to achieve a decent AUROC score. This is because the Weibull distribution used by OpenMax [8] underfits to the handful of support samples for each of the novel test categories occurring in the meta-testing phase, which in turn renders it ineffective in recalibrating the softmax classification scores of the underlying Proto-Net. Since ReFOCS does not rely solely on the softmax scores for out-of-distribution detection it is able to achieve significantly higher AUROC scores. Using the pre-trained Resnet12 encoder for the meta-training phase boosts ReFOCS's performance even

Fig. 5: **Sample exemplar reconstructions for GTSRB→GTSRB.** Exemplar reconstructions of a few query samples from 1 test episode. (a) Class-wise exemplars provided in the support set. (b) In-distribution query samples. (c) Out-of-distribution query samples. (d) Reconstructed exemplars from the in-distribution queries. (e) Reconstructed Exemplars from the out-of-distribution queries, as hypothesized when out-of-distribution samples are fed into ReFOCS it fails to reconstruct the in-distribution class-wise exemplars provided in the support.

further. Pre-training also helps improve the performance of PEELER, although with the Resnet18 backbone, it fails to achieve its reported performance when the results are generated by using its official implementation [1].

In comparison to the variants of SNATCHER [13], our framework is able to achieve comparable performance, even outperforming them in some cases. Since the authors of SNATCHER did not release their code we are able to compare with it for *mini*ImageNet benchmark only, and its results shown in Table 4 are taken directly from their paper [13]. Although SNATCHER is able to effectively use complex architectures like transformers [14] and carefully crafted normalization schemes to improve performance over PEELER, being a thresholding type FSOSR method, it is highly sensitive to the choice of an appropriate threshold. In contrast, ReFOCS eliminates the need for a hand-tuned threshold by utilizing the exemplar reconstruction setup to induce self-awareness in the model, making it more versatile than existing FSOSR baselines.

1. https://github.com/BoLiu-SVCL/meta-open/

4.4 Ablation Studies

In this section, we use two cross-dataset scenarios to perform ablation studies of different components of our framework to understand their contribution toward the final performance.

Impact of using variational encoding.

The reconstruction module of ReFOCS is designed using a VAE [16] due to its better generalization ability [30]. We highlight this by replacing the VAE and experimenting with a standard auto-encoder (AE). As shown in Table 5, both the classification and out-of-distribution detection performance drops significantly. This shows that the improved generalization capability of a VAE is necessary for handling FSOSR scenarios with domain shifts between the training and testing sets.

Impact of weighted prototype computation.

By comparing the second and last row of Table 5 we can see that when our weighted prototype is replaced with a simple one as [11], there is a considerable drop in the classification performance. This shows that utilizing the exemplar information in computing the weighted prototype results in more unbiased prototypical representations of each class, in turn facilitating improved metric-based classification of the in-distribution query samples.

Impact of embedding modulation.

The importance of the embedding modulation can be clearly seen in Table 5. The modulated embedding results in more distinct feature representations with adequate segregation between the seen and unseen classes, thereby making it easier for the novelty module to flag unseen categories. Removing it results in a significant drop in AUROC, and on the other hand, its presence also amplifies the classification performance. This suggests that the improvement of the open class detection is complemented by improvement in seen class categorization.

Input to the Out-of-distribution detector.

We feed a concatenation of the modulated embedding, classification score, and ℓ_2 reconstruction errors to the out-of-distribution detector. Empirically, the combination of all three gives the best results for out-of-distribution detection. We study the impact of removing the classification score or the embedding from the input. As seen from the results in Table 5, in both cases, there is a drop in both the

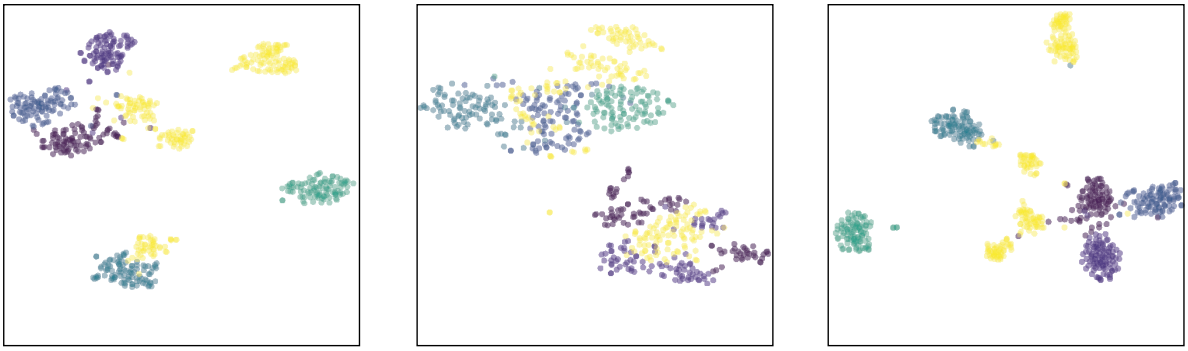

Fig. 6: **t-SNE visualization.** We project the latent space learned via (a) ProtoNet (b) PEELER (c) ReFOCS, on 5 classes of GTSRB→TT100K, on a 2D space using t-SNE. Out-of-distribution queries are in yellow

classification accuracy and the AUROC score. Additionally, we also examine the scenario when we do not use any exemplar for reconstruction and, consequently, feed in just the classification score and the raw embedding as input to the novelty module. We call this variant of ReFOCS, ProtoC+ND, and as seen from Table 5 the removal of the exemplar reconstruction errors results in the biggest drop in performance - with AUROC dropping to 50% in one case (random prediction) - which again consolidates the impact of the exemplars in out-of-distribution detection.

Choice of distance metric.

Computing the logits for classification requires a distance metric to measure the similarity between the prototypes and the sample in the latent space. We experiment with both the cosine and euclidean metrics and choose the cosine distance as it leads to more discriminative embeddings [31], [32] which is particularly important for segregating the open class samples from the support class ones. This is validated by the results shown in Table 6, where we can see that choosing the euclidean distance metric leads to a significant drop in AUROC.

TABLE 6: **Choice of metric.** Effect on performance as the similarity metric is varied. For *mini*Imagenet the Resnet18 [41] backbone is used.

Dataset	Setting	Metric	Acc.	AUROC
GTSRB→TT100K	5-shot	Cosine	**83.36±0.76**	**85.25±0.56**
		Euclidean	78.77±0.74	81.42±0.51
Belga→Flickr32	1-shot	Cosine	**66.2±1.02**	**72.9±0.83**
		Euclidean	48.73±0.99	50.61±0.95
*mini*ImageNet	5-shot	Cosine	**79.06±1.17**	**69.91±0.51**
		Euclidean	**79.09±1.15**	52.83±0.48

Out-of-distribution performance with varying Openness

Openness as defined in [43] is shown below,

$$openness = 1 - \sqrt{\frac{2N_{train}}{N_{test} + N_{target}}} \tag{11}$$

TABLE 7: Comparison of open recognition performance under different openness. Results are in terms of F1 score.

	0%	8.7%	18.4%	23.3%	29.3%
	GTSRB→TT100K				
PEELER	59.11	51.06	48.91	43.66	40.28
ReFOCS	**68.07**	**62.74**	**57.93**	**56.29**	**52.31**
	Belga→Flickr32				
PEELER	48.91	46.73	41.22	39.95	37.14
ReFOCS	**53.65**	**50.52**	**46.19**	**43.78**	**41.26**

where N_{train} is the number of known classes seen during training, N_{test} is the number of classes that will be observed during testing, and N_{target} is the number of open classes to be recognized during testing [43]. For an N-way K-shot problem, the support and the training set are the same, and therefore, $N_{train} = N_{test} = N$. N_{target} refers to the number of open classes we sample $\mathbb{Q}_{out}$. We vary N_{target} from 5 to 15 and observe the open-set recognition performance in terms of averaged F1 score [43] for both ReFOCS and PEELER. As validated by Table 7 ReFOCS consistently outperforms the thresholding-based PEELER on all the levels of openness.

Embedding Visualization.

In Fig. 6, we compare the t-SNE [44] visualization of the embedding spaces induced by all the competing methods. We can see from Fig. 6 that, in general, ReFOCS achieves more distinct class clusters in comparison to both PEELER [12] and ProtoNet [11], with adequate segregation between seen and unseen classes.

5 Conclusion

In this work, we present a novel strategy for addressing few-shot open-set recognition. We frame the few-shot open-set classification task as a meta-learning problem similar to [12], but unlike their strategy, we do not solely rely on thresholding softmax scores to indicate the openness of a sample. We argue that existing thresholding type FSOSR methods [12], [13] rely heavily on the choice of a carefully

tuned threshold to achieve good performance. Additionally, the proclivity of softmax to overfit to unseen classes makes it an unreliable choice as an open-set indicator, especially when there is a dearth of samples. Instead, we propose to use a reconstruction of exemplar images as a key signal to detect out-of-distribution samples. The learned embedding which is used to classify the sample is further modulated to ensure a proficient gap between the seen and unseen class clusters in the feature space. Finally, the modulated embedding, the softmax score, and the quality reconstructed exemplar are jointly utilized to cognize if the sample is in-distribution or out-of-distribution. The enhanced performance of our framework is verified empirically over a wide variety of few-shot tasks and the results establish it as the new state-of-the-art. In the future, we would like to extend this approach to more cross-domain few-shot tasks, including videos.

6 Acknowledgement

This work was partially supported by US National Science Foundation grant 2008020 and US Office of Naval Research grants N00014-19-1-2264 and N00014-18-1-2252.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[4] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] C. Geng, S.-j. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[6] J. Lu, T. Issaranon, and D. Forsyth, "Safetynet: Detecting and rejecting adversarial examples robustly," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[7] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *Proceedings of International Conference on Learning Representations*, 2017.

[8] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[9] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[10] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 3637–3645.

[11] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4080–4090.

[12] B. Liu, H. Kang, H. Li, G. Hua, and N. Vasconcelos, "Few-shot open-set recognition using meta-learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8795–8804.

[13] M. Jeong, S. Choi, and C. Kim, "Few-shot open-set recognition by transformation consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 566–12 575.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[15] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[16] J. Kim, T.-H. Oh, S. Lee, F. Pan, and I. S. Kweon, "Variational prototyping-encoder: One-shot learning with prototypical images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9462–9470.

[17] M. Rohrbach, S. Ebert, and B. Schiele, "Transfer learning in a transductive setting," in *Advances in neural information processing systems*, 2013, pp. 46–54.

[18] D. S. Raychaudhuri and A. K. Roy-Chowdhury, "Exploiting temporal coherence for self-supervised one-shot video re-identification," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*. Springer, 2020, pp. 258–274.

[19] J. Guan, Z. Lu, T. Xiang, A. Li, A. Zhao, and J.-R. Wen, "Zero and few shot learning with semantic feature synthesis and competitive learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 7, pp. 2510–2523, 2020.

[20] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," 2020.

[21] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *CoRR*, vol. abs/1803.02999, 2018. [Online]. Available: http://arxiv.org/abs/1803.02999

[22] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," 2020.

[23] Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi, "Generative openmax for multi-class open set classification," 2017.

[24] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li, "Open set learning with counterfactual images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[25] S. Kong and D. Ramanan, "Opengan: Open-set recognition via open data generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 813–822.

[26] Y. Wang, B. Li, T. Che, K. Zhou, Z. Liu, and D. Li, "Energy-based open-world uncertainty modeling for confidence calibration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9302–9311.

[27] J. Hwang, S. W. Oh, J.-Y. Lee, and B. Han, "Exemplar-based open-set panoptic segmentation network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1175–1184.

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

[29] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[30] B. Dai, Y. Wang, J. Aston, G. Hua, and D. Wipf, "Connections with robust pca and the role of emergent sparsity in variational autoencoder models," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1573–1614, 2018.

[31] J. Liu, L. Song, and Y. Qin, "Prototype rectification for few-shot learning," *ArXiv*, vol. abs/1911.10713, 2019.

[32] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4367–4375.

[33] N. Savinov, A. Raichuk, R. Marinier, D. Vincent, M. Pollefeys, T. Lillicrap, and S. Gelly, "Episodic curiosity through reachability," in *International Conference on Learning Representations (ICLR)*, 2019.

[34] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural networks : the official journal of the International Neural Network Society*, vol. 32, pp. 323–32, 02 2012.

[35] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2110–2118.

[36] A. Joly and O. Buisson, "Logo retrieval with a contrario visual query expansion," in *Proceedings of the 17th ACM international conference on Multimedia*, 2009, pp. 581–584.

[37] P. Letessier, O. Buisson, and A. Joly, "Scalable mining of small visual objects," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 599–608.

[38] S. Romberg, L. G. Pueyo, R. Lienhart, and R. Van Zwol, "Scalable logo recognition in real-world images," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, 2011, pp. 1–8.

[39] H. Su, X. Zhu, and S. Gong, "Deep learning logo detection with data expansion by synthesising context," *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 530–539, 2017.

[40] H.-Y. Tseng, H.-Y. Lee, J.-B. Huang, and M.-H. Yang, "Cross-domain few-shot classification via learned feature-wise transformation," *arXiv preprint arXiv:2001.08735*, 2020.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[43] X. Sun, Z. Yang, C. Zhang, K.-V. Ling, and G. Peng, "Conditional gaussian distribution learning for open set recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 480–13 489.

[44] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

Sayak Nag received his Bachelor's degree in Instrumentation and Electronics Engineering engineering from Jadavpur University, Kolkata, India. Currently, he is pursuing a Ph.D. in the Department of Electrical and Computer Engineering at the University of California, Riverside. His broad research interests include computer vision and machine learning with a focus on few-shot learning, meta-learning, open-set recognition, and weakly-supervised learning.

Dripta S. Raychaudhuri received his Ph.D. in Electrical and Computer Engineering from the University of California, Riverside, and his Bachelor's degree in Electrical and Telecommunication engineering from Jadavpur University, Kolkata, India. He is currently an Applied Scientist at Amazon AWS, USA. His broad research interests include computer vision and machine learning with a focus on multi-task learning, domain adaptation, and imitation learning.

Sujoy Paul received his PhD in Electrical and Computer Engineering from the University of California, Riverside, and his Bachelor's degree in Electronics and Telecommunication Engineering from Jadavpur University. He is currently a Research Scientist at Google Research, India. His broad research interest includes Computer Vision and Machine Learning, focusing on semantic segmentation, human action recognition, domain adaptation, weak supervision, active learning, reinforcement learning, and so on.

Amit K. Roy-Chowdhury received his PhD from the University of Maryland, College Park (UMCP) in 2002 and joined the University of California, Riverside (UCR) in 2004 where he is a Professor and Bourns Family Faculty Fellow of Electrical and Computer Engineering, Director of the Center for Robotics and Intelligent Systems, and Cooperating Faculty in the department of Computer Science and Engineering. He leads the Video Computing Group at UCR, working on foundational principles of computer vision, image processing, and statistical learning, with applications in cyber-physical, autonomous, and intelligent systems. He has published over 200 papers in peer-reviewed journals and conferences. He has also published two monographs on camera networks and wide-area tracking. He is on the editorial boards of major journals and program committees of the main conferences in his area. He is a Fellow of the IEEE and IAPR, received the Doctoral Dissertation Advising/Mentoring Award 2019 from UCR, and the ECE Distinguished Alumni Award from UMCP.

Reconstruction guided Meta-learning for Few Shot Open Set Recognition (Supplementary material)

1 DATASETS

In this section we go over the details of each dataset used in our experiments. A summary of the statistics of each dataset is shown in Table 1.

1.1 Traffic Sign Datasets.

GTSRB. This dataset [**?**] is one of the largest traffic-sign dataset. It is comprised of 43 classes broadly categorized under prohibitory, danger and mandatory categories. For GTSRB→GTSRB, the training set contains a total of 39,209 images and the test set has a total of 12,630 images. These images have variations in illumination, shading as well as, resolution [**?**].

TT100K. The Tsinghua-Tencent 100K (TT100K) [**?**] is a Chinese traffic sign detection dataset. This dataset has more than 200 categories, out of which we used the ones with valid annotation and a corresponding exemplar. Similar to [**?**] this amounts to a total of 36 valid classes to work with.

1.2 Brand Logo Datasets

BelgaLogos. BelgaLogos [**?**], [**?**] is comprised of $10,000$ images of everyday brand logos commonly found in almost every aspect of daily life. The images have significant perturbations such as blurring, lighting and contrast variations as well as, occlusions. The dataset is also riddled with significant class imbalance as it contains classes with as little as 2 samples making it suitable for only 1-shot learning problems. Similar to [**?**] we collect $9,475$ images from BelgaLogos categorised among 37 logo classes to construct our logo recognition dataset. The images of this dataset have a lot more variations compared to the traffic sign datasets, especially blurring and occlusion, which in-turn makes learning harder.

FlickrLogos-32. This dataset [**?**] is comprised of a collection of images corresponding to the 32 brand logos of Flickr. We use the splits introduced in [**?**], which is comprised of a total of $3,372$ cropped logo images.

TopLogo-10. This dataset [**?**] contains logo images from 10 brands related to popular cloth, shoes, and accessory brands. The logo images are obtained from their respective products, and similar to [**?**] the collected images were cropped from bounding box annotations and divided among 11 classes where the 'Adidas' class is divided into 'Adidas-logo' and the 'Adidas-text'.

1.3 Natural Image Dataset

***mini*ImageNet.** Originally proposed by Vinyals et. al. [**?**], *mini*ImageNet has become a benchmark dataset for few-shot classification [**?**], [**?**]. It is derived from the larger ILSVRC-12 dataset [**?**] and is comprised of 100 classes each having 600 colored images, which are popularly resized to 84×84 [**?**], [**?**]. We use the same splits as [**?**], which comprises a training set with 64 classes, a validation set with 16 classes, and a test set with 20 classes.

TABLE 1: **Number of samples and classes in each Dataset.**

	GTSRB	TT100K	BelgaLogos	Flickr-32	TopLogos	*mini*ImageNet
Number of Samples	$51,839$	$11,988$	$9,585$	$3,404$	848	$60,000$
Total Classes	43	36	37	32	11	100

2 ADDITIONAL IMPLEMENTATION DETAILS

2.1 Setup details

For the traffic sign and brand logo recognition, the learning rate remains constant throughout the meta-training stage. For experiments on *mini*ImageNet we drop the learning rate by a factor of 10 after every 20000 episodes.

2.2 Training details

For the traffic sign and brand logo classification, all modules of ReFOCS are meta-trained end-to-end. However, for *mini*ImageNet training the entire framework end-to-end results in over-powering of the reconstruction module resulting in decreased classification accuracy. We suspect this problem is due to the difficult image-to-exemplar translation task for *mini*ImageNet which is brought about by the fact that the estimated exemplars of *mini*ImageNet tend to have much more variability compared to those of the traffic sign and brand logo datasets. In order to rectify this, we first meta-train the VAE encoder with just the cross-entropy classification loss (Eq. 6), following which the weights of the trained encoder are fixed and utilized to jointly train the decoder and the novelty detection module using a combination of the reconstruction loss (Eq. 2) and the novelty detection loss (Eq. 8).

For *mini*ImageNet we also explore starting the meta-training phase by using the pre-trained encoder (also called backbone) when using Resnet-12. For both backbones the pre-trained encoder is directly used for the exemplar estimation process explained in section 3.5. For the Resnet12 case, the pre-training is done following [**?**] and during meta-training, the encoder's weights are initialized with those of the pre-trained encoder.

3 More Ablations

TABLE 2: Comparison of self-reconstruction and exemplar reconstruction guided meta-training for FSOSR.

Experiment	GTSRB→TT100K 5-way 5-shot		Belga→Flickr32 5-way 1-shot	
	Acc.(%)	AUROC(%)	Acc.(%)	AUROC(%)
Self-Reconstruction	41.73 ± 0.86	49.92 ± 0.67	33.96 ± 1.13	48.55 ± 0.91
Estimated Exemplar Reconstruction	80.09 ± 0.79	79.48 ± 0.62	65.15 ± 1.02	69.08 ± 0.84
Canonical Exemplar Reconstruction	83.36 ± 0.76	85.25 ± 0.56	66.29 ± 1.02	72.98 ± 0.83

TABLE 3: Comparison of different distance metrics for exemplar reconstruction and modulation factor. For the current setup, the distance metrics proposed in Eq 10 and Eq 7 are used.

Experiment	GTSRB→TT100K 5-way 5-shot		Belga→Flickr32 5-way 1-shot	
	Acc.(%)	AUROC(%)	Acc.(%)	AUROC(%)
Case 1	79.84 ± 0.77	**79.38 ± 0.63**	**65.08 ± 1.06**	**69.21 ± 0.84**
Case 2	**79.95 ± 0.80**	75.89 ± 0.62	**65.02 ± 1.01**	66.95 ± 0.88
Case 3	79.43 ± 0.79	75.34 ± 0.68	64.91 ± 1.04	66.87 ± 0.85
Current Setup	**80.09 ± 0.79**	**79.48 ± 0.62**	**65.15 ± 1.02**	**69.08 ± 0.84**

3.1 Why exemplar reconstruction is important for FSOSR?

As we have mentioned before, in few-shot learning, naively applying self-reconstruction (reconstructing the input query sample itself) will not enable the learning of an effective task-irrelevant representation. Therefore, it is necessary to anchor the samples of each class to a high-quality abstraction of that class, which we achieve using class-specific exemplars. To verify this, we compare self-reconstruction with exemplar reconstruction for the FSOSR tasks of GTSRB→TT100K and Belga→Flickr32. As observed from Table 2, it can be seen that naively applying self-reconstruction leads to a catastrophic failure to generalize to FSOSR tasks as evident from the lower than random AUROC values as well as the significantly lower classification accuracies.

3.2 Choice of metric for exemplar estimation and modulation factor κ

For the nearest neighbor based exemplar estimation (Eq 10) and the modulation factor κ (Eq 7) we choose the ℓ_2 norm (euclidean distance) and the ℓ_1 norm (manhattan distance), respectively. One might ask why a cosine similarity-based distance such as the following,

$$Distance_{cosine}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}^T\mathbf{y}}{||\mathbf{x}||_2||\mathbf{y}||_2} \tag{1}$$

was not used for these two. This is because, for the exemplar estimation (Eq 10), we observe that the simple euclidean distance suffices to get the best exemplar. However, for the embedding modulation factor κ, it is necessary to use the ℓ_1 norm. This is because if a cosine similarity-based distance is used for κ, it is constrained between 0 and 1, whereas utilizing the ℓ_1 norm allows κ to have ≥ 1 values enabling greater scaling down of the embedding of out-of-distribution samples. To study this, we experiment by considering the following three cases,

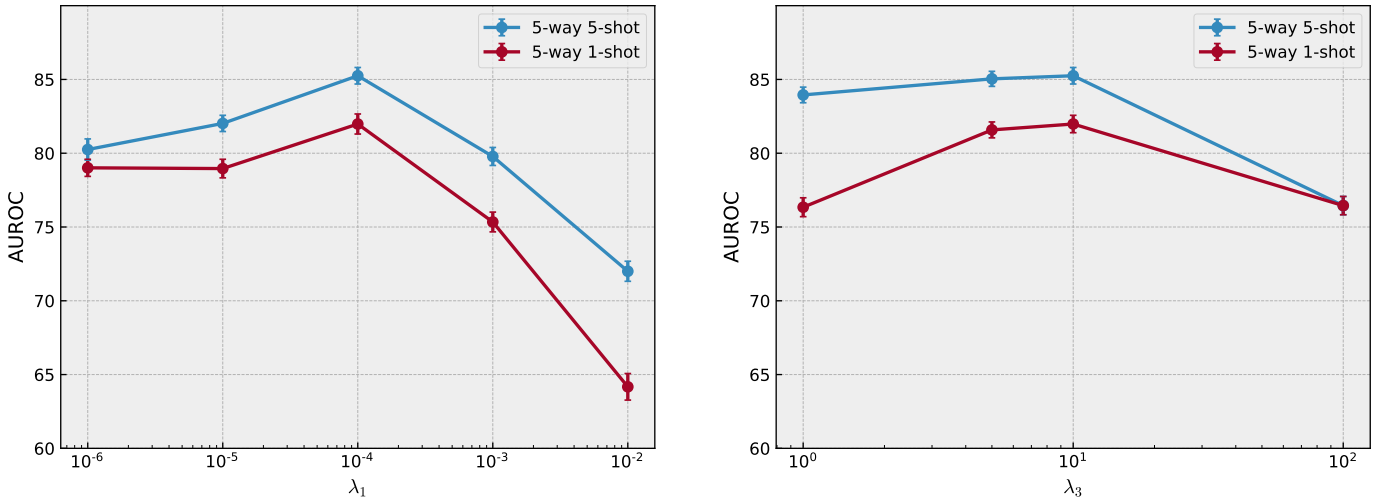

Fig. 1: **Hyperparameter Analysis**. (a) λ_2 & λ_3 are fixed at 5 and 10, changing only the reconstruction loss term. (b) λ_1 and λ_2 are fixed at 10^{-4} and 10 and λ_3 is varied.

- **Case 1**: In Eq 10 we use $Distance_{cosine}()$ as shown above and keep κ the same as Eq 7.
- **Case 2**: For κ we use $Distance_{cosine}()$ and keep the exemplar estimation same as Eq 10.
- **Case 3**: For both κ and exemplar estimation we use $Distance_{cosine}()$.

We compare the results obtained from these three cases with the current setup of Eq 7 and Eq 10. From the observations tabulated in Table 3, we can see that estimating the exemplars using the cosine distance, as opposed to the euclidean distance, does not significantly impact the Accuracy or AUROC. Therefore, choosing the euclidean distance suffices. On the other hand, if $Distance_{cosine}()$ is used to compute κ (Case 2 and Case 3), we observe a significant drop in AUROC since the scaling of the query embedding becomes limited.

3.3 Hyperparameter variation

Fig. 1 shows how the open-set AUROC is affected by the hyperparameters λ_1 or λ_3 for the GTSRB→TT100K task. In both cases, the classification term λ_2 is fixed at 10. In Fig. 1a, when λ_1 is very low, the open-set detection is hampered due to poor quality of the reconstruction and when it is increased beyond 10^{-4} reconstruction becomes the sole objective of the model, thus the open-set detection again degrades. From Fig. 1b we can see that for both the 5-shot and 1-shot cases increasing λ_3 causes the AUROC to improve the knee point of $\lambda_3 = 10$, after which it starts to degrade.

4 More Qualitative Visualizations

Some exemplar reconstructions for the *mini*ImageNet → *mini*ImageNet experiment are shown in Fig 2. Since the exemplars are also natural images, the high variation in lighting and backgrounds makes the reconstruction much harder. However, ReFOCS performs the way it's intended to, and as such out-of-distribution samples fail to reconstruct the in-distribution exemplars.

Fig. 2: **Sample exemplar reconstructions for *mini*ImageNet→*mini*ImageNet.** Exemplar reconstructions of a few query samples from one test episode. (a) Class-wise exemplars are provided in the support set. (b) In-distribution query samples. (c) Out-of-distribution query samples. (d) Reconstructed exemplars from the in-distribution queries. (e) Reconstructed Exemplars from the out-of-distribution queries, as hypothesized when out-of-distribution samples are fed into ReFOCS it fails to reconstruct the in-distribution class-wise exemplars provided in the support.