HELA-VFA: A Hellinger Distance-Attention-based Feature Aggregation Network for Few-Shot Classification

Gao Yu Lee, Tanmoy Dam, Daniel Puiu Poenar, Vu N. Duong
Nanyang Technological University (NTU), Singapore
{GAOYU001, tanmoy.dam, EPDPuiu, vu.duong}@ntu.edu.sg

Md Meftahul Ferdaus
University of New Orleans, USA
ferdaus57@gmail.com

Abstract

Enabling effective learning using only a few presented examples is a crucial but difficult computer vision objective. Few-shot learning have been proposed to address the challenges, and more recently variational inference-based approaches are incorporated to enhance few-shot classification performances. However, the current dominant strategy utilized the Kullback-Leibler (KL) divergences to find the log marginal likelihood of the target class distribution, while neglecting the possibility of other probabilistic comparative measures, as well as the possibility of incorporating attention in the feature extraction stages, which can increase the effectiveness of the few-shot model. To this end, we proposed the HELlinger-Attention Variational Feature Aggregation network (HELA-VFA), which utilized the Hellinger distance along with attention in the encoder to fulfill the aforementioned gaps. We show that our approach enables the derivation of an alternate form of the lower bound commonly presented in prior works, thus making the variational optimization feasible and be trained on the same footing in a given setting. Extensive experiments performed on four benchmarked few-shot classification datasets demonstrated the feasibility and superiority of our approach relative to the State-Of-The-Arts (SOTAs) approaches.

1. Introduction

Computer vision has achieved numerous breakthroughs in recent decades, and can be attributed mainly to deep learning along with the availability of huge data for effective supervised learning. However, this is still a far cry from our human visual and brain system's ability to learn and recognize new objects and classes by virtue from being exposed to only its few instances. Few-Shot Learning (FSL) has been recently introduced and leveraged in attempts to mimic the aforementioned recognition capability and has found promising results relative to the standard CNN-based supervised learning networks in benchmark datasets. Therefore, interest among machine and deep learning practitioners and researchers has been on the rise. An example of a recent review of FSL can be found at [1].

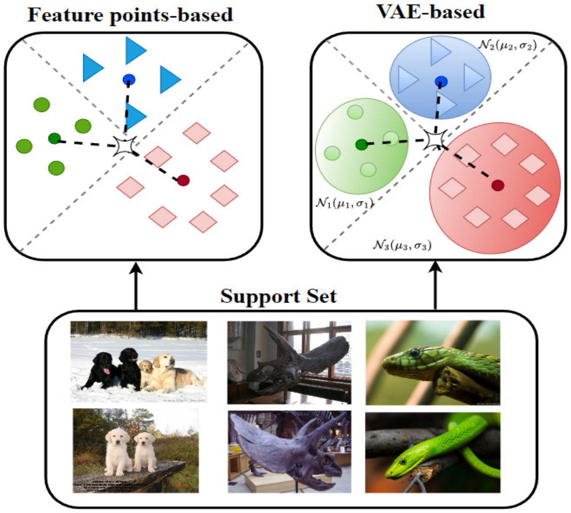
Feature points-based
VAE-based
$\mathcal{N}_2(\mu_2, \sigma_2)$
$\mathcal{N}_1(\mu_1, \sigma_1)$
$\mathcal{N}_3(\mu_3, \sigma_3)$
Support Set

Figure 1. The differences between point-based and VAE-based FSL feature processing in the embedded space, using a 3-class classification scheme as an example. The support set images consisted of images belonging to the same class (column) and are extracted from the miniImageNet. The dotted lines represents the Euclidean distance between the sample feature points and the centroid of each class prototype.

Prior works in FSL involved learning the degree of similarity between images presented to the model and defining a distance measure in which similar objects are assigned the same class label, while vastly different objects are assigned a different class label. All these can be performed either directly (e.g., Siamese network [2]), or by embedding the visual features into a higher-dimensional space (e.g., prototypical network [3]). For the latter, this results in embedded feature points of the same class clustered in the same region, and the regions can be segmented and represented pictorially as on the left of Figure 1. Specifically, the prototypical network utilized a point estimate approach, which although has been shown to yield State-Of-the-Art (SOTA) performance in many benchmark datasets, still has its limitations [4]. The first is that the point estimations may be inaccurate when there are limited support sample points unevenly distributed in the embedded space. This is supported by [5] which emphasized that the quality of the support sets is a crucial factor and it can be difficult to estimate the mean of a class via only a few samples. The second is that a single embedding of the feature point is insufficient to indicate a class and may lack interpretability. The third is that if there is any data distribution inconsistency between the training and the test phase, overfitting can easily occur.

To address these limitations, the aforementioned work proposed a distribution-based estimation approach which reduces sample biases and account for intra-class variances in a better manner than that of point estimation approaches. A pictorial representation is shown on the right of Fig.1. Still, an intractable integral needs to be computed if one opts for a probabilistic computation, and hence a variational approximation that involves minimizing the Kullback-Leibler (KL) divergence (D_{KL}) [6] was invoked. Currently, most, if not all, existing image-based variational few-shot approaches invoked D_{KL}, while we have noted the existence of other possible classes of divergences (in which D_{KL} is a subset of) that are very less explored, but which can potentially be utilized and compared with the D_{KL} baseline. The Hellinger distance (D_H) [7] is one such example. It is the probabilistic analog of the Euclidean distance, and make computations and comparisons more straightforward and easier than the D_{KL}. Furthermore, one does not have to worry about its value diverging to infinity as one of the input probability distribution goes to zero during a computational workflow. Lastly, there are existing works on utilizing the Hellinger distance for classification in other domains ([8], [9], [10]) and, as far as we know, there are little to no works on incorporating such distance in FSL classification on benchmark datasets.

Another emerging aspect is the incorporation of attention in the feature extractor stage of the few-shot model. Introducing attention allows the model to focus on essential visual features and omit irrelevant ones, which increases the model's efficiency. Various attention mechanisms have been increasingly introduced to FSL approaches, which include dual attention [11], self-attention [12], and attention with weight fusion [13]. These works has also demonstrated in their ablation studies that incorporating attention leads to improvement in the classification accuracy scores, as compared to when such mechanisms were omitted. We also noted the existence of variational-based FSL that incorporates attention (e.g., [14] for AE-GAN with self-attention). However, the latter's design focused on addressing some limits of KL divergence-based VAE (e.g.,the retention of image border features). All these considerations motivate us to explore a VAE with an alternate divergence measure with attention and observe how it could enhance the baseline variational FSL classification performances.

Therefore, in this paper, we introduce a HELlinger-Attention Variational Feature Aggregation (HELA-VFA) network that performs variational inference via D_H, and improves upon the VFA-Net via introducing attention during the feature extraction pipeline. While contrastive loss ($\ell_{contrastive}$) and categorical cross-entropy (ℓ_{CCE}) are commonly utilized during the training of the FSL, we additionally introduced the Hellinger similarity softmax loss (ℓ_{Hess}), which replaces the cosine similarity in the simCLR [15] softmax loss with the Hellinger distance, and serves as the Euclidean distance-like variation of the loss function but in a probability distribution space. We combined our proposed loss with the other relevant loss functions commonly utilized in variational inference-based FSL to enhance our network's training performance. Overall, our contributions to this work are as follows:

- We proposed the HELA-VFA, which pioneered a Hellinger distance-based feature aggregation scheme with attention mechanism imbued in the feature extraction stages. We show that in spite of a difference in probability distribution distance measures, we can still derive our form of the Evidence Lower BOund (ELBO) and cast the optimization problem in a similar manner as that of KL divergences.

- We deployed the designated ℓ_{Hess}, amalgamating it linearly with the traditionally employed loss function in variational inference-based FSL methodologies, like ℓ_{CCE} and the reconstruction loss ℓ_{rec}. This integration serves to augment the training performance of our network.

- Empirical simulations were conducted on four widely recognized FSL benchmark datasets: FC-100 [16], CIFAR-FS [17], miniImageNet [18], and tieredImageNet [19]. The results illustrate the practicability and supremacy of our method in comparison to the current SOTA FSL approaches.

2. Related Works

One of the first variational-inference based FSL is presented by [4] which highlights the limitations and benefits of approaching a distributive-based distance measure in the embedded space instead of feature points as in many prototypical-based FSL approaches. A variational few-shot feature aggregation approach was proposed by [5] which fused the support and query features during the batch training, which is then utilized for few-shot object detection.

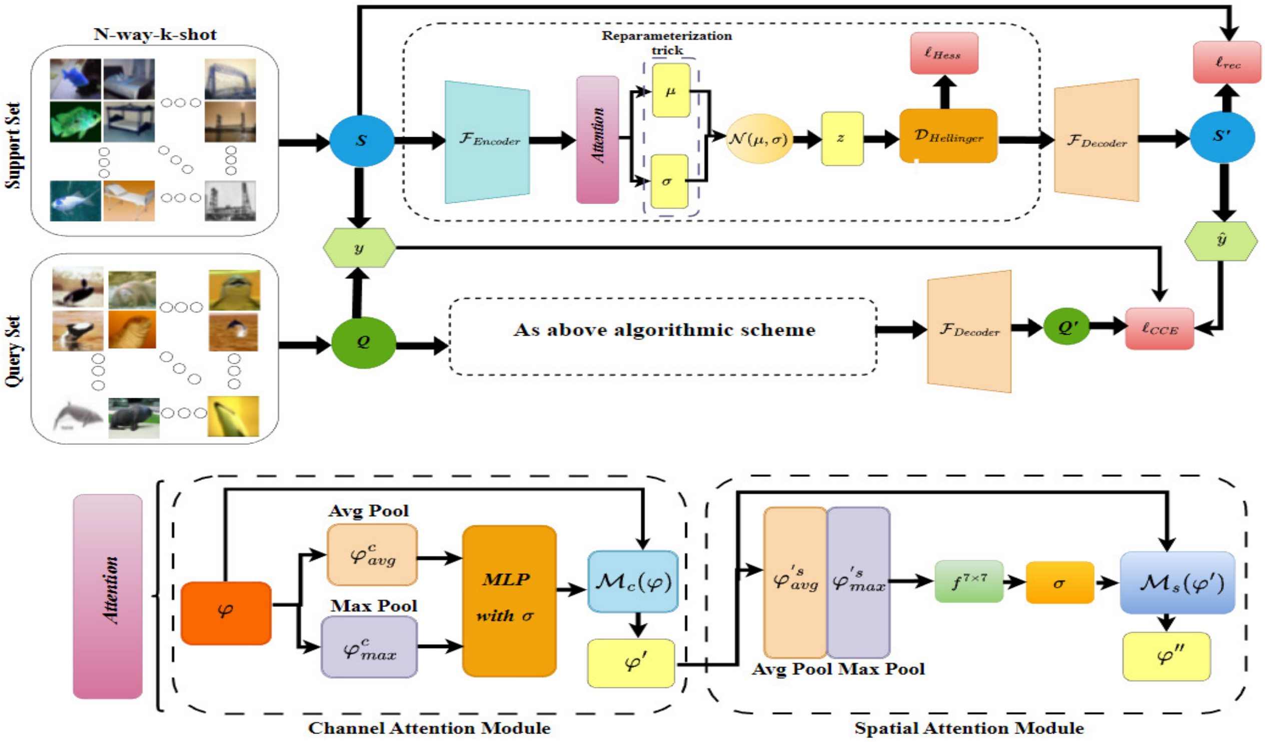

Figure 2. HELA-VFA algorithmic architecture (**top**) and the attention mechanism architecture (**bottom**). In the top diagram, S and S' denotes the original and reconstructed images respectively , while Q and Q' denotes the corresponding quantity but for the query set. $\hat{y}$ and y denotes the predicted label after training and ground truth label respectively. The rest of the quantities are introduced along in section 3. The above network is allows a general a N-way-k-shot training and evaluation.

This encourages inter-class interactions which promotes class-agnostic representations and reduces the confusion between the base and novel classes in the prediction phase. The transductive decoupled variational inference network (TRIDENT) was introduced by [20] which decoupled the image representation into semantic and label latent variables, and performed inference via fusing the informative in the support and query set, similar to the feature aggregation work. However, unlike the two aforementioned studies, the TRIDENT incorporated attention in the feature extraction phase. We realized that there also exists attention-based FSL works but in the aspect of query point embedding, such as [21]. However, to the best of our knowledge, TRIDENT is one of the first work that incorporated both attention and variational inference simultaneously, and that it demonstrated the feasibility of combining both components in the same network to enhance the classification performances. However, all of the methods described approached the log marginal likelihood optimization via associating the link between ELBO and the KL divergence. The KL divergences is a subset of the *f-divergences* which also include the Hellinger distance. Unlike the KL divergences which used the ratio of the logarithm between two probability distributions, the Hellinger distance manifests as the square of the differences between the square root of the distributions. Apart from the advantages of such distance measure as mentioned in the introductory section, the corresponding link between the ELBO and the Hellinger distance can be derived, meaning that the variational optimization process can be performed on the same footing as that of the prior related works regardless of the datasets used.

3. Our Approach

We illustrate the algorithmic architecture of our HELA-VFA at the top of Figure 2. In summary, the key additions in our network include the attention mechanism in the encoder, the usage of the Hellinger distance, and the introduction of our ℓ_{Hess} along with the standard losses utilized in related previous works.

3.1. Attention Mechanism

The attention mechanism is incorporated in our network's feature encoder and involved the channel and spatial attention modules. For the channel module, the spatial dimension of the extracted intermediate feature maps φ) are processed using both average $({\varphi^c}_{avg})$ and max pooling ($\boldsymbol{\varphi^c}_{\boldsymbol{max}}$). The output features are then inputted into a shared Multi-Layer Perceptron (*MLP*), with the resultant features $\boldsymbol{\mathcal{M}_c}(\boldsymbol{\varphi})$ combined using an element-wise summation. Mathematically speaking,

$$\boldsymbol{\mathcal{M}_c}(\boldsymbol{\varphi}) = \sigma(MLP(\boldsymbol{\varphi^c}_{\boldsymbol{avg}}) + MLP(\boldsymbol{\varphi^c}_{\boldsymbol{max}})). \tag{1}$$

where σ is the sigmoid activation function. For the spatial module, the features obtained from both pooling in the spatial domain, $\boldsymbol{\varphi^s}_{\boldsymbol{avg}}$ and $\boldsymbol{\varphi^s}_{\boldsymbol{max}}$, are concatenated in the channel axis. A 2D convolutional layer with a filter size of 7×7 ($f^{7\times7}$) is applied to compute the spatial attention map $\boldsymbol{\mathcal{M}_s}(\boldsymbol{\varphi})$.

$$\boldsymbol{\mathcal{M}_s}(\boldsymbol{\varphi'}) = \sigma(f^{7\times7}([\boldsymbol{\varphi'^s}_{\boldsymbol{avg}}; \boldsymbol{\varphi'^s}_{\boldsymbol{max}}])). \tag{2}$$

Combining our attention modules,

$$\begin{aligned} \boldsymbol{\varphi'} &= \boldsymbol{\mathcal{M}_c}(\boldsymbol{\varphi}) \otimes \boldsymbol{\varphi}, \\ \boldsymbol{\varphi''} &= \boldsymbol{\mathcal{M}_s}(\boldsymbol{\varphi'}) \otimes \boldsymbol{\varphi'}. \end{aligned} \tag{3}$$

$\otimes$ represents the outer product between the extracted features ($\boldsymbol{\varphi'}$, $\boldsymbol{\varphi''}$) and the attention maps. The algorithmic structure of our attention mechanism is illustrated at the bottom of Figure 2.

3.2. HELA-VFA Optimization

The goal of variational inference in our HELA-VFA is to minimize the divergence measure, which is equivalent to maximizing the Evidence Lower Bound (ELBO). The bound obtained for the previous works involved the KL divergence D_{KL}. Here we will show that such bound also exist, and can also be obtained via the Hellinger distance (D_H). We start with the square of its definition:

$$D_H^2 = 1 - \int \left(\sqrt{p_\theta(z|\mathcal{T})q_\phi(z|\mathcal{S})}\right) dz, \tag{4}$$

where $p_\theta(z|\mathcal{T})$ denotes the prior distribution of z with our knowledge of $\mathcal{T}$, and $q_\phi(z|\mathcal{S})$ represents the true parameterized distribution conditioned on $\mathcal{S}$. Swapping the positions and taking the logarithm of both sides,

$$\log \int \left(\sqrt{p_\theta(z|\mathcal{T})q_\phi(z|\mathcal{S})}\right) dz = \log\left(1 - D_H^2\right). \tag{5}$$

The mathematical definition of the ELBO is

$$ELBO = \int q_\phi(z|\mathcal{S})\log\left(\frac{p_\theta(\mathcal{T}, z)}{q_\phi(z|\mathcal{S})}\right) dz. \tag{6}$$

The derivation of the bounds in terms of D_H is long and hence **its proof is provided in the supplementary material**. Here we simply state the result, which serves as our HELA-VFA optimization condition:

$$ELBO' = \int \log(p_\theta(\mathcal{T}))dz + \log\left(1 - D_H^2\right)^2 \tag{7}$$

where $\int \log(p_\theta(\mathcal{T}))dz$ is the *evidence* term, and the new ELBO term, $ELBO'$ is written as $\frac{ELBO}{q_\phi(z|\mathcal{S})}$.

For comparison, we included the ELBO for model utilizing the KL divergence:

$$ELBO = \int \log(p_\theta(\mathcal{T}))dz - D_{KL}(q_\phi(z|\mathcal{S})||p_\theta(z|\mathcal{T})), \tag{8}$$

and we can see that the differences lie in that for our approach, not only the square of the logarithm of the difference in D_H^2 with respect to 1 is needed, which ensured that the range of D_H^2, and hence D_H lies in $[0, 1]$, our $ELBO'$ is manifested as the linear sum of the evidence and the logarithm of the square of the parenthesis containing the D_H, rather than the linear differences between the evidence and the relevant divergence (as in the case of D_{KL}). However, note that $\log\left(1 - D_H^2\right)^2$ took on the range $(-\infty, 0]$, while $D_{KL}(q_\phi(z|\mathcal{S})||p_\theta(z|\mathcal{T}))$ took on the range $[0, \infty)$, and when both terms approaches zero (which occurs when $q_\phi(z|\mathcal{S}) = p_\theta(z|\mathcal{T})$), the respective ELBO values equate to the evidence. Hence maximizing the evidence term is equivalent to maximizing the respective lower bounds.

Similar to the KL divergence, we can express D_H in terms of the means and standard deviations of the i$^{\text{th}}$ class prototypes in the prior distribution (μ_i, σ_i) and j$^{\text{th}}$ posterior distribution (μ_j, σ_j). The mathematical form is of a Hellinger distance between two Gaussian distribution, which is derived in [22], and we simply state the form here:

$$D_H^2 = 1 - \sqrt{\frac{2\sigma_i\sigma_j}{\sigma_i^2 + \sigma_j^2}}\exp\left(-\frac{1}{4}\frac{(\mu_i - \mu_j)^2}{(\sigma_i^2 + \sigma_j^2)}\right). \tag{9}$$

3.3. HELA-VFA Model Training

We utilized the ResNet-12 [23] encoder backbone for the feature extraction. Regardless of the divergence measure used, in a N-way FSL scheme, the support and target set are partitioned into N subset each of a certain class (i.e., $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_N$). Therefore there are N posterior class-specific distribution required for estimation, all conditioned on $\mathcal{S}$. These distributions satisfies a Gaussian with a diagonal covariance structure $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$, where the mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ are a set of class distributed values (i.e., $\boldsymbol{\mu} = \{\boldsymbol{\mu}_i | i = 1, 2, ...N\}$, and $\boldsymbol{\sigma} = \{\boldsymbol{\sigma}_i | i =$

$1, 2, ...N\}$). The sampling process from these distribution parameterized by our model to obtain the latent variable z is not a differentiable process, which makes any gradient-based learning processes not feasible. To overcome this, we employed the reparameterization trick on the latent variable distribution following [24], i.e., $z = \mu + \sigma \odot \epsilon$, where ϵ is generated from a standard normal distribution $\mathcal{N}(0, 1)$, and $\odot$ denotes the Hadamard product.

Once the specific distributions are estimated, the probability of the target data samples belonging to their respective classes are calculated and the maximum values obtained serves as the predicted label, which is compared to the ground-truth label y_i via a categorical cross entropy loss ℓ_{CCE} [25]

$$\ell_{CCE} = -\sum_{i=1}^{Q} y_i \log(p(\hat{y}_i = y_i | \mathcal{T}_{meta})) \tag{10}$$

where Q is the number of query samples in a batch-trained meta task $\mathcal{T}_{meta}$.

In addition to ℓ_{CCE}, we introduced the Hellinger Similarity softmax loss function ℓ_{Hess}, which is motivated by the cosine similarity softmax loss utilized in contrastive similarity FSL approaches, but replaces the similarity in the latter with the Hellinger similarity. The motivation is that the cosine similarity can be thought of as the dot product between the class-relevant feature vector and the class prototype, and since the Hellinger distance can be thought of as the probabilistic analog of the vector computation in a Euclidean distance-like manner, we can first compute the class-specific distributions of the query vectors $v_{(Q,i)}$ along with the prototype vectors before computing its probability of belonging to the k$^{\text{th}}$-class and applying the loss. Mathematically speaking, our ℓ_{Hess} manifests as

$$\ell_{Hess} = -\sum_{i=1}^{N_t} y_i \log(p(y = j | v_{(Q,i)})) \tag{11}$$

where N_t denotes the number of training samples, and $p(y = j | v_{(Q,i)})$ is computed from

$$p(y = j | v_{(Q,i)}) = \frac{\exp(-\{v_{(Q,i)}, c_{(Q,j)}\})}{\sum_{i=1}^{N} \exp\left(-\{v_{(Q,i)}, c_{(Q,j)}\}\right)} \tag{12}$$

in which $c_{(Q,j)}$ denotes the jth-class prototype from query set. Note that in the denominator, the sum is taken over the total number of classes N. The Hellinger similarity is hence contained in the term $\{v_{(Q,i)}, c_{(Q,j)}\}$ in the parenthesis of $p(y = j | v_{(Q,i)})$ and consequently, ℓ_{Hess}.

The other loss required in our network is the reconstruction loss ℓ_{rec}, also utilized in prior related works, which aids in image reconstruction via the decoder $\mathcal{F}_{dec}$ in our HELA-VFA, and manifests as the L1-norm between the input and the reconstructed support image, S and S' respectively.

$$\ell_{rec} = ||\mathcal{F}_{dec}(z) - S|| = ||S' - S||. \tag{13}$$

Overall, the loss function for our HELA-VFA network training, ℓ_{HELA} is as follows:

$$\ell_{HELA} = \ell_{CCE} + \lambda_1 \cdot \ell_{Hess} + \lambda_2 \cdot \ell_{rec}, \tag{14}$$

where λ_1, λ_2 represents the weight parameters of ℓ_{Hess} and ℓ_{rec} with respect to ℓ_{CCE}.

4. Experiments

We evaluated our approach relative to the State-Of-The-Arts (SOTAs) FSL approaches on the FC-100, CIFAR-FS, miniImageNet, and tieredImageNet. The selected SOTAs for each dataset is the same as that laid out in [26].

The FC-100 dataset is a subset of the CIFAR-100 and comprises 60 meta-training classes, 20 meta-validation classes, and 20 meta-testing classes. Each train-valid-test split contains 600 images, each of size 32 × 32. The small image size corresponds to low image quality, and hence make few-shot learning a challenge. The CIFAR-FS dataset is also a subset of CIFAR-100 and contains 64 meta-training classes, 16 meta-validation classes, and 20 meta-testing classes. The image size is also of 32 × 32, and the high intra-class similarity among the images also makes the few-shot tasks challenging.

The miniImageNet dataset comprises a wide variety of images from 100 classes, and each class has 600 images. The image size is set to 84 × 84, and the train-valid-test split comprises images of 64, 16, and 20 classes, respectively. Finally, the tieredImageNet, like the miniImageNet, is a subset of the ImageNet in which 351 classes are utilized for meta-training, 97 classes are for meta-validation, and 160 classes are for meta-testing.

All simulations were performed using the Google Colab's Tesla A100, V100 and T4 Graphical Processing Units (GPU), with PyTorch as the underlying libraries. In line with numerous FSL literatures, we reported the few-shot classification accuracy (in %) for our method, adopting the N-way-1-shot and N-way-5-shot strategies, where N is the number of classes utilized for the evaluation. Regarding the training of our network, the Adam optimization algorithm was employed, featuring a learning rate of 1e-3. Furthermore, the training incorporated an epoch count of 200, weight decay parameter of 1e-5, and a batch size comprising of 32 instances.

5. Results and Discussions

Table 1, 2, and 3 present the results of the 5-way-1-shot and 5-way-5-shot evaluations for our HELA-VFA approach and its comparison with previous works, specifically on CIFAR-FS, FC-100, miniImageNet, and tieredImageNet

Table 1. The classification accuracy (in %) using the 5-way 1-shot and 5-way 5-shot learning evaluation for our HELA-VFA relative to the SOTAs on the **CIFAR-FS**.

Methods	**5-way-1-shot**	**5-way-5-shot**
ProtoNet [3]	72.7±0.7	83.5±0.5
MetaOptNet [27]	72.6±0.7	84.3±0.5
DSN-MR [28]	75.6±0.9	86.2±0.6
RFS-Simple [29]	71.5±0.8	86.0±0.5
RFS-distill [29]	73.9±0.8	86.9±0.5
IER-distill [30]	77.6±1.0	89.7±0.6
PAL [31]	77.1±0.7	88.0±0.5
SKD-Gen1 [32]	76.6±0.9	88.6±0.5
Label Halluc [33]	78.0±1.0	89.4±0.6
FeLMi [26]	78.2±0.7	89.5±0.5
HELA-VFA	**78.9±0.4**	**90.7±0.7**

All CNN backbone used is the ResNet-12.

Table 2. The classification accuracy (in %) using the 5-way 1-shot and 5-way 5-shot learning evaluation for our HELA-VFA relative to the SOTAs on the **FC-100**.

Methods	**5-way-1-shot**	**5-way-5-shot**
ProtoNet	37.5±0.6	52.5±0.6
MetaOptNet	41.1±0.6	55.5±0.6
TADAM [16]	40.1±0.4	56.1±0.4
MTL [34]	45.1±1.8	57.6±0.9
RFS-Simple	42.6±0.7	59.1±0.6
Deep-EMD [35]	46.5±0.8	63.2±0.7
RFS-Simple	42.6±0.7	59.1±0.6
RFS-distill	44.6±0.7	60.9±0.6
IER-distill	48.1±0.8	65.0±0.7
PAL	47.2±0.6	64.0±0.6
SKD-Gen1	46.5 ±0.8	64.2±0.8
AssoAlign [36]	45.8 ±0.5	59.7±0.6
InfoPatch [37]	43.8 ±0.4	58.0±0.4
Label Halluc	47.3±0.7	67.9±0.7
FeLMi	49.0±0.7	68.7±0.7
HELA-VFA	**50.3±0.3**	**69.1±0.2**

All CNN backbone used is the ResNet-12, except for AssoAlign which utilized a ResNet-18 backbone.

datasets. With the exception of AssoAlign, all the chosen techniques employed ResNet-12 as their encoder backbone.

One notable observation is that for all the methods, the accuracy derived from the 5-way-5-shot evaluations invariably surpasses that of the 5-way-1-shot evaluations. This can be attributed to the heightened challenge posed by using few-shot learning when there are fewer examples per training batch, given that the support set offers less supplementary information to guide the model towards accurate predictions. A second point of note is that for both 5-way-1-shot and 5-way-5-shot evaluations, the classification accuracies procured by all assessed methodologies are lower for FC-100 compared to CIFAR-FS. This indicates that the FC-100 poses a more significant challenge for Few-Shot Learning (FSL) methods to achieve accurate classifications than CIFAR-FS. As per the accuracy figures tabulated for miniImageNet and tieredImageNet, the difficulty in executing few-shot classification falls between the aforementioned two datasets. Among these, miniImageNet is slightly more challenging than tieredImageNet. Our HELA-VFA method consistently excels over other state-of-the-art approaches across almost all datasets, albeit with marginal enhancement. For example, in the CIFAR-FS evaluation, our HELA-VFA model improves upon FeLMi (which holds the second-best classification accuracy across all outlined datasets, barring tieredImageNet in the 5-way-1-shot approach) by 0.90% in the 5-way-1-shot evaluation, and by 1.34% in the 5-way-5-shot evaluation. Similarly, for FC-100, our approach once again surpasses FeLMi, this time by 2.65% and 0.58% in the 5-way-1-shot and 5-way-5-shot evaluations, respectively. In the miniImageNet analysis, our method exhibited better classification performances than FeLMi by 1.04% and 0.70% in the 5-way-1-shot and 5-way-5-shot evaluations, respectively. Lastly, for tieredImageNet, our method outperformed FeLMi by 1.26% and 0.57% in the 5-way-1-shot and 5-way-5-shot evaluations, respectively. Nonetheless, it is critical to mention that our approach's performance remains marginally lower than that of IER-distill in the 5-way-1-shot evaluation for tieredImageNet.

5.1. Ablation Studies

We conducted an ablation study to analyze the role of each novel components in our HELA-VFA on the classification performances, and reported the results for each shot setting in Table 4 for miniImageNet and 5 for FC-100 (although we have experimented and noted similar trends for the remaining two datasets). The configurations attempted are illustrated in the tables, with the listed third and fourth configurations involved replacing ℓ_{Hess} in ℓ_{HELA} with ℓ_{KL} to compare the role of D_{KL} and D_H in the classification performances. We can observe from the tables that relatively lower values are reported for all shot setting when attention was not incorporated, regardless of whether ℓ_{KL} or ℓ_{Hess} is used (or when both are not used). This quantitatively emphasized the role of attention in the feature extraction stage. However, we also observed that the obtained values utilizing ℓ_{KL} were lower than that of ℓ_{Hess}, regardless of whether attention is incorporated or not. This can be attributed to D_{KL} not being an actual distance metric (since it does not satisfied the triangle inequality), thus for a pair of not-so-distant distribution configurations in which the value are closely similar to one another, (p, q), (p', q') where $p = q'$ and $q = p'$, ℓ_{KL} may yield different values which affects our network's feature clustering computation.

Table 3. The classification accuracy (in %) using the 5-way 1-shot and 5-way 5-shot learning evaluation for our HELA-VFA relative to the SOTAs on the **miniImageNet** and **tieredImageNet**.

	Datasets			
	miniImageNet		**tieredImageNet**	
Methods	**5-way-1-shot**	**5-way-5-shot**	**5-way-1-shot**	**5-way-5-shot**
ProtoNet	60.4±0.8	78.0±0.6	65.7±0.9	83.4±0.7
MetaOptNet	62.6±0.6	78.6±0.5	66.0±0.7	81.6±0.5
MTL	61.2±1.8	75.5±0.8	65.6±1.8	80.6±0.9
TADAM	58.5±0.3	76.7±0.3	-	-
Shot-Free [38]	59.0±0.4	77.6±0.4	66.9±0.4	82.6±0.4
Deep-EMD	65.9±0.8	82.4±0.6	71.2±0.9	86.0±0.6
FEAT [39]	66.8±0.2	82.1±0.1	70.8±0.2	84.8±0.2
DSN-MR	64.6±0.7	79.5±0.5	67.4±0.8	82.9±0.6
Neg-Cosine [40]	63.9±0.8	81.6±0.6	-	-
P-Transfer [41]	64.2±0.8	80.4± 0.6	-	-
MELR [42]	67.4±0.4	83.4±0.3	72.1±0.5	87.0±0.4
TapNet [43]	61.7±0.2	76.4±0.1	63.1±0.2	80.3±0.1
IEPT [44]	67.1±0.4	82.9±0.3	72.2±0.5	86.7±0.3
RFS-Simple	62.0±0.6	79.6±0.4	69.7±0.7	84.4±0.6
RFS-distill	64.8±0.8	82.4±0.4	71.5±0.7	86.0±0.5
IER-distill	66.9±0.8	84.5±0.5	**72.7±0.9**	86.6±0.8
SKD-Gen1	66.5±1.0	83.2±0.5	72.4 ±1.2	86.0±0.6
AssoAlign	60.0±0.7	80.4±0.7	69.3±0.6	86.0±0.5
Label Halluc	67.0±0.7	85.9±0.5	72.0±0.9	86.8±0.6
FeLMi	67.5±0.8	86.1±0.4	71.6±0.9	87.1±0.6
HELA-VFA	**68.2±0.3**	**86.7±0.7**	72.5±0.5	**87.6±0.1**

All CNN backbone used is the ResNet-12, except for AssoAlign which utilized a ResNet-18 backbone.

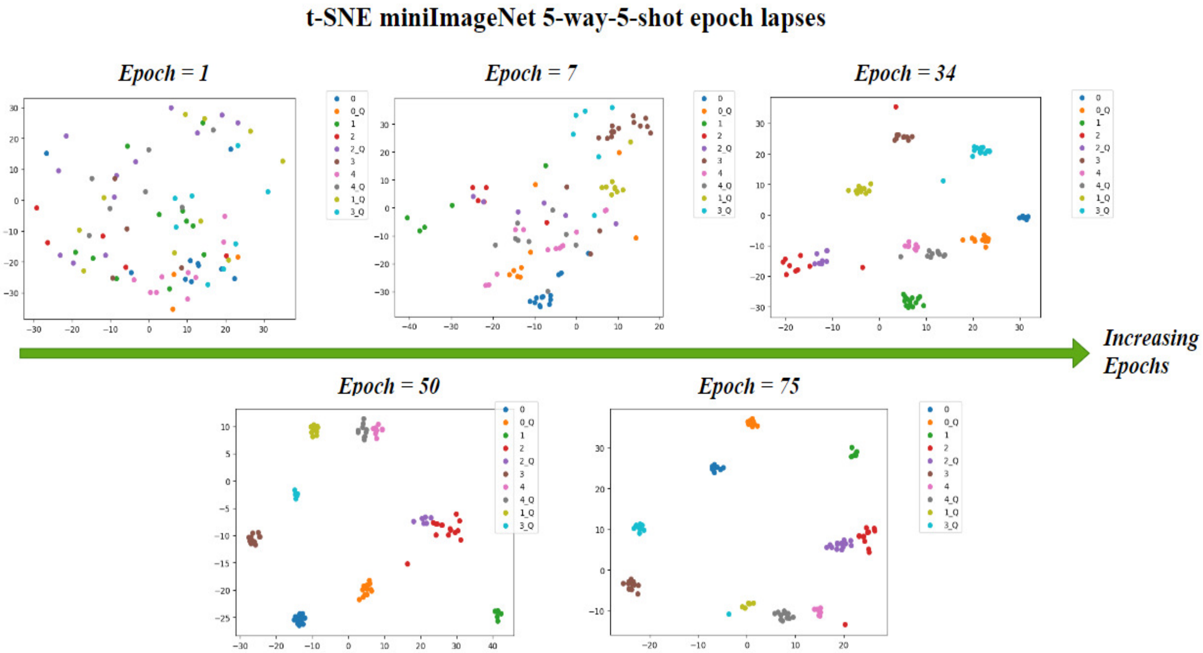

Figure 3. The t-SNE plots as a function of epochs for the 5-way-5-shot miniImageNet scenario. The respective support label (0-4) and its corresponding query label (with a Q after the number label) are also summarized in the legend (**zoom in to better see the labels**).

Table 4. Ablation study results on the HELA-VFA for the various configurations tabulated below on the **miniImageNet** dataset.

Configuration	5-way-1-shot	5-way-5-shot
Attention + ℓ_{Hess}	**68.2±0.3**	**86.7±0.7**
No Attention + ℓ_{Hess}	63.3±0.5	82.6±1.0
Attention + ℓ_{KL}	61.5±0.6	75.8±0.8
No Attention + ℓ_{KL}	58.9±0.9	71.7±0.5
W/o ℓ_{Hess} or ℓ_{KL}	57.5±0.9	73.4±0.4
W/o Att, ℓ_{Hess} or ℓ_{KL}	56.7±0.3	72.6±0.6

Table 5. Ablation study results on the HELA-VFA for the various configurations tabulated below on the **FC-100** dataset.

Configuration	5-way-1-shot	5-way-5-shot
Attention + ℓ_{Hess}	**50.3±0.3**	**69.1±0.2**
No Attention + ℓ_{Hess}	46.2±1.2	65.9±0.9
Attention + ℓ_{KL}	45.6±0.5	65.1±0.6
No Attention + ℓ_{KL}	43.5±1.1	63.5±0.7
W/o ℓ_{Hess} or ℓ_{KL}	40.7±0.4	47.6±1.1
W/o Att, ℓ_{Hess} or ℓ_{KL}	37.6±1.0	42.0±0.8

When we mentioned ℓ_{KL} as above, we mean replacing the ℓ_{Hess} with ℓ_{KL} in the ℓ_{HELA}.

5.2. t-SNE analysis

We also offered an examination of our model's ability to differentiate latent feature variables from respective test classes, examined as a function of t-SNE epochs, by using the miniImageNet dataset as a representative example. This analysis is visually depicted in Figure 3, with the epochs selected for scrutiny being 1, 7, 34, 50, and 75. Query labels within the dataset's testing set are designated as $0-4$, while the labels corresponding to the support set are denoted as 0_Q-4_Q. To facilitate a more straightforward illustration, the depicted plots correspond to the 5-way-5-shot scenarios; however, it should be noted that comparable trends can be extrapolated from the other configuration results.

Upon analysis, it becomes evident that the test set plots exhibit a progressive degree of class cluster segmentation as the epochs increase. This trend mirrors the improvement in our model's classification accuracy throughout the t-SNE runs. More specifically, it is observed that the query set labels generally become more proximate to the clusters of the support set labels. The $(2, 2\_Q)$ set, highlighted in red and purple respectively, and the $(4, 4\_Q)$ set, depicted in pink and gray respectively, are found to be the initial sets to achieve such clustering, a trend that is clearly observable from epoch 34 onwards. This pattern is then followed by the $(0, 0\_Q)$ set, highlighted in dark blue and orange, which demonstrates evidence of clustering from epoch 50 and beyond. Subsequently, the $(3, 3\_Q)$ set exhibits a similar clustering pattern from epoch 75 onwards. Notably, the $(1, 1\_Q)$ set presents a distinctive challenge in terms of clustering, with the latent variable distance displaying fluctuations across the epochs. Despite this, a gradual narrowing of these distances has been observed as the epochs advance. The latent feature plots presented in epoch 1 serve to reinforce earlier assertions regarding the high intra-class similarity present within the data, as the majority of the latent features are closely entwined prior to the learning process. Undoubtedly, these prominent challenges elucidate the complexities inherent in the selected dataset and the resultant obstacles to achieving effective few-shot learning. Despite these challenges, the methodological approach we have employed has delivered notable success, surpassing the accuracy values recorded by existing state-of-the-art models, and achieving the desired clustering of the support and query set subsequent to the training process, as is corroborated by the epoch lapses displayed in Figure 3.

This victory, however, should be interpreted in the correct context. While our method has indeed exceeded previous benchmarks, it must be acknowledged that the margin of surpassing is not overwhelmingly extensive. The majority of the values obtained by our methodology do not register an improvement exceeding 5%. This proportion serves as a reminder of the inherent intricacies associated with few-shot learning and the perennial need for further advancements in this field. It simultaneously stands as a testament to the method's potential, illustrating its capacity to perform commendably even within the constraints of this challenging arena. This nuanced understanding of our method's achievements underscores both the opportunities for and the constraints on significant advancement within the field of few-shot learning. The ongoing challenge is to continue to refine and improve these techniques to consistently yield even greater improvements.

6. Conclusions

We put forth HELA-VFA, an innovative solution implementing variational inference-based few-shot classification. Our method employs the Hellinger distance metric, along with an attention mechanism embedded in the network's encoder, to extract and aggregate pivotal features from both the support and query sets. HELA-VFA distinguishes itself as one of the first methodologies to incorporate attention and variational inference within few-shot learning, thus offering a new avenue for the incorporation of alternative probability distribution-based distance metrics in comparative FSL studies. Our methodology was subjected to rigorous testing on four of the aforementioned prominent few-shot benchmark datasets, with our approach exceeding the classification performances of the FSL SOTAs in both the 5-way-1-shot and 5-way-5-shot evaluations. This underlines the feasibility and efficacy of our algorithmic design within the benchmark context.

References

[1] G. Y. Lee, T. Dam, M. M. Ferdaus, D. P. Poenar, and V. N. Duong, "Unlocking the capabilities of explainable fewshot learning in remote sensing," *arXiv preprint arXiv:2310.08619*, 2023. 1

[2] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, no. 1. Lille, 2015. 1

[3] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017. 1, 6

[4] J. Zhang, C. Zhao, B. Ni, M. Xu, and X. Yang, "Variational few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1685–1694. 2

[5] J. Han, Y. Ren, J. Ding, K. Yan, and G.-S. Xia, "Few-shot object detection via variational feature aggregation," *arXiv preprint arXiv:2301.13411*, 2023. 2

[6] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951. 2

[7] E. Hellinger, "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen." *Journal für die reine und angewandte Mathematik*, vol. 1909, no. 136, pp. 210–271, 1909. 2

[8] J. Grzyb, J. Klikowski, and M. Woźniak, "Hellinger distance weighted ensemble for imbalanced data stream classification," *Journal of Computational Science*, vol. 51, p. 101314, 2021. 2

[9] A. Kumari and U. Thakar, "Hellinger distance based oversampling method to solve multi-class imbalance problem," in *2017 7th International Conference on Communication Systems and Network Technologies (CSNT)*. IEEE, 2017, pp. 137–141. 2

[10] Z. Z. Al-Shamaa, S. Kurnaz, A. D. Duru, N. Peppa, A. H. Mirnezami, Z. Z. Hamady *et al.*, "The use of hellinger distance undersampling model to improve the classification of disease class in imbalanced medical datasets," *Applied bionics and biomechanics*, vol. 2020, 2020. 2

[11] Z. Jiang, B. Kang, K. Zhou, and J. Feng, "Few-shot classification via adaptive attention," *arXiv preprint arXiv:2008.02465*, 2020. 2

[12] X. Huang and S. H. Choi, "Sapenet: Self-attention based prototype enhancement network for few-shot learning," *Pattern Recognition*, vol. 135, p. 109170, 2023. 2

[13] X. Meng, X. Wang, S. Yin, and H. Li, "Few-shot image classification algorithm based on attention mechanism and weight fusion," *Journal of Engineering and Applied Science*, vol. 70, no. 1, pp. 1–14, 2023. 2

[14] L. Lin, X. Liu, and W. Liang, "Improving variational auto-encoder with self-attention and mutual information for image generation," in *Proceedings of the 3rd international conference on video and image processing*, 2019, pp. 162–167. 2

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607. 2

[16] B. Oreshkin, P. Rodríguez López, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," *Advances in neural information processing systems*, vol. 31, 2018. 2, 6

[17] L. Bertinetto, J. F. Henriques, P. H. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," *arXiv preprint arXiv:1805.08136*, 2018. 2

[18] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016. 2

[19] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," *arXiv preprint arXiv:1803.00676*, 2018. 2

[20] A. Singh and H. Jamali-Rad, "Transductive decoupled variational inference for few-shot classification," *arXiv preprint arXiv:2208.10559*, 2022. 3

[21] Q.-H. Nguyen, C. Q. Nguyen, D. D. Le, H. H. Pham, and M. N. Do, "Enhancing few-shot image classification with cosine transformer," *arXiv preprint arXiv:2211.06828*, 2022. 3

[22] L. Pardo, *Statistical inference based on divergence measures*. CRC press, 2018. 4

[23] 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. 4

[24] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," *Advances in neural information processing systems*, vol. 28, 2015. 5

[25] H. Cheng, Y. Wang, H. Li, A. C. Kot, and B. Wen, "Disentangled feature representation for few-shot image classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 5

[26] A. Roy, A. Shah, K. Shah, P. Dhar, A. Cherian, and R. Chellappa, "Felmi: Few shot learning with hard mixup," in *Advances in Neural Information Processing Systems*, 2022. 5, 6

[27] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 657–10 665. 6

[28] G. Tao, L. Weichao, H. Yanmin, and L. Yu, "Graph-based prototypical network for few-shot learning," in *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE, 2021, pp. 234–237. 6

[29] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: a good embedding is all you need?" in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 266–282. 6

[30] M. N. Rizve, S. Khan, F. S. Khan, and M. Shah, "Exploring complementary strengths of invariant and equivariant representations for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 836–10 846. 6

[31] J. Ma, H. Xie, G. Han, S.-F. Chang, A. Galstyan, and W. Abd-Almageed, "Partner-assisted learning for few-shot image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 573–10 582. 6

[32] J. Rajasegaran, S. Khan, M. Hayat, F. S. Khan, and M. Shah, "Self-supervised knowledge distillation for few-shot learning," *arXiv preprint arXiv:2006.09785*, 2020. 6

[33] Y. Jian and L. Torresani, "Label hallucination for few-shot classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 6, 2022, pp. 7005–7014. 6

[34] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 403–412. 6

[35] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 203–12 213. 6

[36] A. Afrasiyabi, J.-F. Lalonde, and C. Gagné, "Associative alignment for few-shot image classification," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 18–35. 6

[37] Y. Gao, N. Fei, G. Liu, Z. Lu, and T. Xiang, "Contrastive prototype learning with augmented embeddings for few-shot learning," in *Uncertainty in Artificial Intelligence*. PMLR, 2021, pp. 140–150. 6

[38] A. Ravichandran, R. Bhotika, and S. Soatto, "Few-shot learning with embedded class models and shot-free meta training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 331–339. 7

[39] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8808–8817. 7

[40] B. Liu, Y. Cao, Y. Lin, Q. Li, Z. Zhang, M. Long, and H. Hu, "Negative margin matters: Understanding margin in few-shot classification," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 438–455. 7

[41] Z. Shen, Z. Liu, J. Qin, M. Savvides, and K.-T. Cheng, "Partial is better than all: revisiting fine-tuning strategy for few-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, 2021, pp. 9594–9602. 7

[42] N. Fei, Z. Lu, T. Xiang, and S. Huang, "Melr: Meta-learning via modeling episode-level relationships for few-shot learning," in *International Conference on Learning Representations*, 2021. 7

[43] S. W. Yoon, J. Seo, and J. Moon, "Tapnet: Neural network augmented with task-adaptive projection for few-shot learning," in *International conference on machine learning*. PMLR, 2019, pp. 7115–7123. 7

[44] M. Zhang, J. Zhang, Z. Lu, T. Xiang, M. Ding, and S. Huang, “Iept: Instance-level and episode-level pretext tasks for few-shot learning,” in *International Conference on Learning Representations*, 2021. 7