Adversarial Feature Hallucination Networks for Few-Shot Learning

Kai Li[1], Yulun Zhang[1], Kunpeng Li[1], Yun Fu[1,2]

[1]Department of Electrical and Computer Engineering, Northeastern University, Boston, USA
[2]Khoury College of Computer Science, Northeastern University, Boston, USA

{kaili,kunpengli,yunfu}@ece.neu.edu, yulun100@gmail.com

Abstract

The recent flourish of deep learning in various tasks is largely accredited to the rich and accessible labeled data. Nonetheless, massive supervision remains a luxury for many real applications, boosting great interest in label-scarce techniques such as few-shot learning (FSL), which aims to learn concept of new classes with a few labeled samples. A natural approach to FSL is data augmentation and many recent works have proved the feasibility by proposing various data synthesis models. However, these models fail to well secure the discriminability and diversity of the synthesized data and thus often produce undesirable results. In this paper, we propose Adversarial Feature Hallucination Networks (AFHN) which is based on conditional Wasserstein Generative Adversarial networks (cWGAN) and hallucinates diverse and discriminative features conditioned on the few labeled samples. Two novel regularizers, i.e., the classification regularizer and the anti-collapse regularizer, are incorporated into AFHN to encourage discriminability and diversity of the synthesized features, respectively. Ablation study verifies the effectiveness of the proposed cW-GAN based feature hallucination framework and the proposed regularizers. Comparative results on three common benchmark datasets substantiate the superiority of AFHN to existing data augmentation based FSL approaches and other state-of-the-art ones.

1. Introduction

The rich and accessible labeled data fuel the revolutionary success of deep learning [7, 46, 20]. However, in many specific real applications, only limited labeled data are available. This motivates the investigation of few-shot learning (FSL) where we need to learn concept of new classes based on a few labeled samples. To combat with deficiency of labeled data, some FSL methods resort to enhance the discriminability of the feature representations such that a simple linear classifier learned from a few labeled samples can reach satisfactory classification re-

sults [39, 36, 38]. Another category of methods investigate techniques of quickly and effectively updating a deep neural network with a few labeled data, either by learning a meta-network and the corresponding updating rules [9, 24, 32, 28], or by learning a meta-learner model that generates some components of a classification network directly from the labeled samples [21, 12, 34]. Alternatively, the third group of methods address this problem with data augmentation by distorting the labeled images or synthesizing new images/features based on the labeled ones [4, 10, 35, 5].

Our proposed method falls into the data augmentation based category. The basic assumption of approaches in this category is that the intra-class cross-sample relationship learned from seen (training) classes can be applied to unseen (test) classes. Once the cross-sample relationship is modeled and learned from seen classes, it can be applied on the few labeled samples of unseen class to hallucinated new ones. It is believed that the augmented samples can diversify the intra-class variance and thus help reach sharper classification boundaries [45]. Whatever data augmentation technique is used, it is critical to secure discriminability of the augmented samples, as otherwise they shall cast catastrophic impact on the classifier. On the other hand, the decision boundary of a classifier can be determined precisely only when labeled samples exhibit sufficient intra-class variance. Thus, diversity of the augmented samples is also of a crucial role. This is in fact the essential motivation of investigating data augmentation for FSL, as a few labeled samples encapsulate limited intra-class variance.

Though various data augmentation based FSL methods have been proposed recently, they fail to simultaneously guarantee discriminability and diversity of the synthesized samples. Some methods learn a finite set of transformation mappings between samples in each base (label-rich) classes and directly apply them to seed samples of novel (label-scarce) classes. However, the arbitrary mapping may destroy discriminability of the synthesized samples [6, 15, 35]. Other methods synthesize samples specifically for certain tasks which regularize the synthesis process [41, 28]. Thus, these methods can guarantee discriminability of the synthe-

sized samples. But the task would constrain the synthesis process and consequently the synthesized samples tend to collapse into certain modes, thus failing to secure diversity.

To avoid limitations of the existing methods, we propose Adversarial Feature Hallucination Networks (AFHN) which consists of a novel conditional Wasserstein Generative Adversarial Networks (cWGAN) [13] based feature synthesis framework and two novel regularizers. Unlike many other data augmentation based FSL approaches that perform data augmentation in the image space [3, 6, 4], our cWGAN based framework hallucinates new features by using the features of the seed labeled samples as the conditional context. To secure discriminability of the synthesized features, AFHN incorporates a novel classification regularizer that constrains *the synthesized features being of high correlation with features of real samples from the same class while of low correlation with those from the different classes*. With this constraint, the generator is encouraged to generate features encapsulating discriminative information of the class used as the conditional context.

It is more complicated to ensure diversity of the synthesized features, as conditional GANs are notoriously susceptible to the mode collapse problem that only samples from limited distribution modes are synthesized. This is caused by the use of usually high dimensional and structured data as the condition tends to make the generator ignore the latent code, which controls diversity. To avoid this problem, we propose a novel anti-collapse regularizer which assigns high penalty for the case where mode collapse likely occurs. It is derived from the observation that *noise vectors that are closer in the latent code space are more likely to be collapsed into the same mode when mapped to the feature space*. We directly penalize the ratio of the dissimilarity of the two synthesized feature vectors and the dissimilarity of the two noise vectors generating them. With this constraint, the generator is forced to explore minor distribution modes, thus encouraging diversity of the synthesized features.

With discriminative and diverse features synthesized, we can get highly effective classifiers and accordingly appealing recognition results. In summary, the contributions of this paper are as follows: (1) We propose a novel cWGAN based FSL framework which synthesizes fake features by taking those of the few labeled samples as the conditional context. (2) We propose two novel regularizers that guarantee discriminability and diversity of the synthesized features. (3) The proposed method reaches the state-of-the-art performance on three common benchmark datasets.

2. Related Work

Regarding the perspective of addressing FSL, existing algorithms can generally be divided into three categories. The first category of methods aim to enhance the discriminability of the feature representations extracted from im-

ages. To this goal, a number of methods resort to deep metric learning and learn deep embedding models that produce discriminative feature for any given image [33, 39, 36, 38]. The difference lies in the loss functions used. Other methods following this line focus on improving the deep metric learning results by learning a separate similarity metric network [37], task dependent adaptive metric [30], patch-wise similarity weighted metric [14], neural graph based metric [18, 25], etc.

A more common category of algorithms address FSL by enhancing flexibility of a model such that it can be readily updated using a few labeled samples. These methods utilize meta-learning, also called learning to learn, which learns an algorithm (meta-learner) that outputs a model (the learner) that can be applied on a new task when given some information (meta-data) about that task. Following this line, some approaches aim to optimize a meta-learned classification model such that it can be easily fine-tuned using a few labeled data [32, 9, 24, 24, 32, 28, 29]. Other approaches adopt neural network generation and train a meta-learning network which can adaptively generate entire or some components of a classification neural network from a few labeled samples of novel classes [31, 12, 22, 21]. The generated neural network is supposed to be more effective to classify unlabeled samples from the novel classes, as it is generated from the labeled samples and encapsulates discriminative information about these classes.

The last category of methods combat deficiency of the labeled data directly with data augmentation. Some methods try to employ additional samples by some forms of transfer learning from external data [33, 42]. More popular approaches perform data augmentation internally by applying transformations on the labeled images or the corresponding feature representations. Naively distorting the images with common transformation techniques (e.g., adding Gaussian perturbation, color jittering, etc.) is particularly risky as it likely jeopardizes the discriminative content in the images. This is undesirable for FSL as we only have a very limited number of images to be utilized; quality control of the synthesizing results for any single image is crucial as otherwise the classifier could be ruined by the low-quality images. Chen et al. propose a series of methods of performing quality-controlled image distortions by applying perturbation in the semantic feature space [6], shuffling image patches [3] and explicitly learning an image transformation network [4]. Performing data augmentation in the feature space seems more promising as the feature variance directly affects the classifier. Many approaches with this idea have been proposed by hallucinating new samples for novel class based on seen classes [35, 15], composing synthesized representations [5, 44], and using GANs [10, 45].

This paper proposes Adversarial Feature Hallucination Networks (AFHN), a new GAN-based FSL model that aug-

ments labeled samples by synthesizing fake features conditioned on those of the labeled ones. AFHN significantly differs from the two existing GAN based models [45, 10] in the following aspects. First, AFHN builds upon Wasserstein GAN (WGAN) model which is known for more stable performance, while [45, 10] adopt the conventional GAN framework. Second, neither [45] nor [10] has a classification regularizer. The most similar optimization objective in [10] is the one which optimizes the synthesized features as the outlier class (relative to the real class), while that in [45] is a cycle-consistency objective. We instead regularize the synthesized features of being high correlation with real features from the same classes and low correlation with those from the different classes. Third, After training the generator, we learn a standard Softmax classifier using the synthesize features, while [45, 10] utilize them to enhance existing FSL methods. Last, we further propose the novel anti-collapse regularizer to encourage diversity of synthesized features, while [45, 10] do not.

AFHN also bears some similarity with an existing feature hallucination based FSL method [41]. But apparently we adopt the GAN framework which has the discriminator to regularize the features produced by the generator, while [41] uses the simple generative model. Besides, AFHN synthesizes new features to learn a standard Softmax classifier for new classes, while [41] utilizes them to enhance existing FSL classifier. Moreover, we aim to hallucinate diverse features with the novel anti-collapse regularizer, while [41] does not have such an objective.

3. Algorithm

In this section, we first briefly introduce Wasserstein GAN and then elaborate the details of how we build the proposed AFHN model upon it.

3.1. Wasserstein GAN

GAN is a recently proposed generative model that has shown impressive performance on synthesizing realistic images. The generative process in GAN is modeled as a game between two competitive models, the generator and the discriminator. The generator aims to generate from noise fake samples as realistic as possible such that the discriminator cannot tell whether they are real or fake. The discriminator instead tries the best to make the correct judgment. This adversarial game pushes the generator to extensively explores the data distribution and consequently produces more visually appealing samples than conventional generative models. However, it is known that GAN is highly unstable in training. [1] analyzes the convergence properties of the objective function of GAN and proposes the Wasserstein GAN (WGAN) which utilizes the Wasserstein distance in the objective function and is shown of better theoretical properties than the vanilla GAN. We adopt the improved variant

of WGAN [13], which optimizes the following min-max problem,

$$
\min_G \max_D \quad \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathbb{P}_g}[D(\tilde{\mathbf{x}})] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_r}[D(\mathbf{x})] \\
+ \lambda \mathbb{E}_{\hat{\mathbf{x}} \sim \mathbb{P}_{\hat{\mathbf{x}}}}[(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2], \tag{1}
$$

where $\mathbb{P}_r$ is the data distribution and $\mathbb{P}_g$ is the model distribution defined by $\tilde{\mathbf{x}} \sim G(\mathbf{z})$, with $\mathbf{z} \sim p(\mathbf{z})$ randomly sampled from noise distribution p. $\mathbb{P}_{\hat{\mathbf{x}}}$ is defined by sampling uniformly along straight lines between pairs of points sampled from the data distribution $\mathbb{P}_r$ and the generator distribution $\mathbb{P}_g$, i.e., $\hat{\mathbf{x}} = \alpha \mathbf{x} + (1 - \alpha)\tilde{\mathbf{x}}$ with $\alpha \sim U(0, 1)$. The first two terms approximate the Wasserstein distance and the third term penalizes the gradient norm of $\hat{\mathbf{x}}$.

3.2. Adversarial Feature Hallucination Networks

Following the literature, we formally define FSL as follows: Given a distribution of tasks $P(\mathcal{T})$, a sample task $\mathcal{T} \sim P(\mathcal{T})$ is a tuple $\mathcal{T} = (S_{\mathcal{T}}, Q_{\mathcal{T}})$ where the support set $S_{\mathcal{T}} = \{\{\mathbf{x}_{i,j}\}_{i=1}^K, y_j\}_{j=1}^N$ contains K labeled samples from each of the N classes. This is usually known as K-shot N-way classification. $Q_{\mathcal{T}} = \{(\mathbf{x}_q, y_q)\}_{q=1}^Q$ is the query set where the samples come from the same N classes as the support set $S_{\mathcal{T}}$. The learning objective is to minimize the classification prediction risk of $Q_{\mathcal{T}}$, according to $S_{\mathcal{T}}$.

The proposed AFHN approaches this problem by proposing a general conditional WGAN based FSL framework and two novel regularization terms. Figure 1 illustrates the training pipeline.

FSL framework with conditional WGAN. For a typical FSL task $\mathcal{T} = (S_{\mathcal{T}}, Q_{\mathcal{T}})$, the feature extraction network F produces a representation vector for each image. Specifically for an image from the support set $(\mathbf{x}, y) \in S_{\mathcal{T}}$, F generates

$$
\mathbf{s} = F(\mathbf{x}). \tag{2}
$$

When there are multiple samples for class y, i.e., $K > 1$, we simply average the feature vectors and take the averaged vector as the prototype of class y [36]. Conditioned on $\mathbf{s}$, we synthesize fake features for the class.

Unlike previous GAN models which sample a single random noise variable from some distribution, we sample two noise variables $\mathbf{z}_1$ and $\mathbf{z}_2 \sim N(0, 1)$. The generator G synthesizes fake feature $\tilde{\mathbf{s}}_1$ ($\tilde{\mathbf{s}}_2$) taking as input $\mathbf{z}_1$ ($\mathbf{z}_2$) and the class prototype $\mathbf{s}$,

$$
\tilde{\mathbf{s}}_i = G(\mathbf{s}, \mathbf{z}_i), \ i = 1, 2. \tag{3}
$$

The generator G aims to synthesize $\tilde{\mathbf{s}}_i$ to be as similar as possible to $\mathbf{s}$. The discriminator D, taking $\mathbf{z}_i$ and $\mathbf{s}$ as input, tries to discern $\tilde{\mathbf{s}}_i$ as fake and $\mathbf{s}$ as real. Within the WGAN framework, the adversarial training objective is as follows,

$$
L_{GAN_i} = \mathbb{E}_{(\mathbf{x},y) \sim S_{\mathcal{T}}}[D(\tilde{\mathbf{s}}_i, \mathbf{z}_i)] - \mathbb{E}_{(\mathbf{x},y) \sim S_{\mathcal{T}}}[D(\mathbf{s}, \mathbf{z}_i)] \\
+ \lambda \mathbb{E}_{(\mathbf{x},y) \sim S_{\mathcal{T}}}[(\|\nabla_{\tilde{\mathbf{s}}_i} D(\hat{\mathbf{s}}_i, \mathbf{z}_i)\|_2 - 1)^2], \ i = 1, 2. \tag{4}
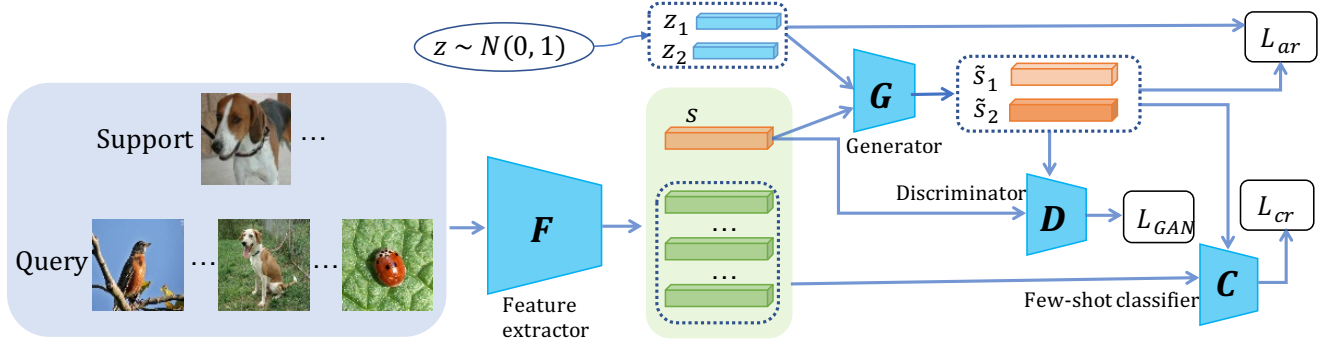$$

Figure 1. Framework of the proposed AFHN. AFHN takes as input a support set and a query set where images in the query set belongs to the sampled classes in the support set. Each image in the support set is fed to the feature extraction network F, resulting the feature embedding $\mathbf{s}$. With $\mathbf{s}$, feature generator G synthesizes two fake features $\tilde{\mathbf{s}}_1$ and $\tilde{\mathbf{s}}_2$, by combining $\mathbf{s}$ with two randomly sampled variables $\mathbf{z}_1$ and $\mathbf{z}_2$. Discriminator D discriminates real feature $\mathbf{s}$ and fake features $\tilde{\mathbf{s}}_1$ and $\tilde{\mathbf{s}}_2$, resulting in the GAN loss L_{GAN}. By analyzing the relationship between $(\mathbf{z}_1, \mathbf{z}_2)$ and $(\tilde{\mathbf{s}}_1, \tilde{\mathbf{s}}_2)$, we get the anti-collapse loss L_{ar}. The proposed few-shot classifier classifies the features of the query images based on the fake features $\tilde{\mathbf{s}}_1$ and $\tilde{\mathbf{s}}_2$. This results in the classification loss L_{cr}.

Simply training the model with the above GAN loss does not guarantee the generated features are well suited for learning a discriminative classifier because it neglects the inter-class competing information among different classes. Moreover, since the conditioned feature vectors are of high dimension and structured, it is likely that the generator will neglect the noise vectors and all synthesized features collapse to a single or few points in the feature space, i.e., the so-called mode collapse problem. To avoid these problems, we append the objective function with a classification regularization term and an anti-collapse regularization term, aiming to encourage both diversity and discriminability of the synthesized features.

Classification regularizer. As our training objective is to classify well samples in the query set $Q_{\mathcal{T}}$, given the support set $S_{\mathcal{T}}$, we encourage discriminability of the synthesized features by requiring them serving well the classification task as the real features. Inspired by [36], we define a non-parametric FSL classifier which calculates the possibility of a query image $(\mathbf{x}_q, y_q) \in Q_{\mathcal{T}}$ of being the same class as synthesized feature $\tilde{\mathbf{s}}_i$ as

$$P(y_q = y|\mathbf{x}_q) = \frac{\exp(\cos(\tilde{\mathbf{s}}_i, \mathbf{q}))}{\sum_{j=1}^{N} \exp(\cos(\tilde{\mathbf{s}}_i^j, \mathbf{q}))}, \quad (5)$$

where $\mathbf{q} = F(\mathbf{x}_q)$. $\tilde{\mathbf{s}}_i^j$ is the synthesized feature for the j-th class and $\cos(\mathbf{a}, \mathbf{b})$ is the Cosine similarity of two vectors. The adoption of Cosine similarity, rather than Euclidean distance as in [36], is inspired by a recent FSL algorithm [12] which proves that Cosine similarity can bound and reduce variance of the features and result in models of better generalization.

With the proposed FSL classifier, the classification regu-

larizer in a typical FSL task is defined as follows:

$$L_{cr_i} = \mathop{\mathbb{E}}_{(\mathbf{x}_q, y_q) \sim Q_{\mathcal{T}}} \left[\frac{1}{N} \sum_{y=1}^{N} y \log[-P(y_q = y|\mathbf{x}_q)] \right], \quad (6)$$

for $i = 1, 2$. We can see that this regularizer explicitly encourages the synthesized features to have high correlation with features from the same class (the conditional context), while low correlation with features from the different classes. To achieve this, the synthesized features must encapsulate discriminative information about the conditioned class and thus secure discriminability.

Anti-collapse regularizer. GAN models are known for suffering from the notorious mode collapse problem, especially conditional GANs where structured and high-dimensional data (e.g., images) are usually used as the conditional contexts. As a consequence, the generator likely ignores the latent code (noises) that accounts for diversity and focuses only on the conditional contexts, which is undesirable. Specifically to our case, our goal is to augment the few labeled samples in the feature space; when mode collapse occurs, all synthesized features may collapse to a single or a few points in the feature space, failing to diversify the labeled samples. Observing that noise vectors that are closer in the latent code space are more likely to be collapsed into the same mode when mapped to the feature space, we directly penalize the ratio of the dissimilarity two synthesized feature vectors and the dissimilarity of the two noise vectors generating them.

Remember that we sample two random variables $\mathbf{z}_1$ and $\mathbf{z}_2$. We generate two fake feature vectors $\tilde{\mathbf{s}}_1$ and $\tilde{\mathbf{s}}_2$ from them. When $\mathbf{z}_1$ and $\mathbf{z}_2$ are closer, $\mathbf{s}_1$ and $\mathbf{s}_2$ are more likely to be collapsed into the same mode. To mitigate this, we

---

Algorithm 1. Proposed FSL algorithm

Input: Training set $\mathcal{D}_t = \{\mathcal{X}_t, \mathcal{Y}_t\}$, parameters λ, α, and β.
Output: Feature extractor F, generator G, discriminator D.

---

1. Train F as a standard classification task using $\mathcal{D}_t$.
while not done **do**
    // Fix G and update D.
    2. Sample from $\mathcal{D}_t$ a batch of FSL tasks $\mathcal{T}_i^d \sim p(\mathcal{D}_t)$.
    For each $\mathcal{T}_i^d$ **do**
        3. Sample a support set $S_\mathcal{T} = \{\{\mathbf{x}_{i,j}\}_{i=1}^K, y_j\}_{j=1}^N$ and
            query set $Q_\mathcal{T} = \{\{\mathbf{x}_{k,j}\}_{k=1}^Q, y_j\}_{j=1}^N$.
        4. Compute prototypes of the N classes $\mathcal{P} = \{\mathbf{s}_j\}_{j=1}^N$,
            where $\mathbf{s}_j = \frac{1}{K} \sum_{i=1}^K F(\mathbf{x}_{i,j})$.
        5. Sample N noise variables $\mathcal{Z}_1 = \{\mathbf{z}_1^j\}_{j=1}^N$ and
            variables $\mathcal{Z}_2 = \{\mathbf{z}_2^j\}_{j=1}^N$.
        6. Generate fake feature sets $\tilde{\mathcal{Z}}_1 = \{\tilde{\mathbf{z}}_1^j\}_{j=1}^N$
            and $\tilde{\mathcal{Z}}_2 = \{\tilde{\mathbf{z}}_2^j\}_{j=1}^N$ according to Eq. (3).
        7. Update D by maximizing Eq. (8).
    end For
    // Fix D and update G.
    8. Sample from $\mathcal{D}_t$ a batch of FSL tasks $\mathcal{T}_i^g \sim p(\mathcal{D}_t)$.
    For each $\mathcal{T}_i^g$ **do**
        9. Execute steps 3 - 7.
        10. Update G by minimizing Eq. (8).
    end For
end while

---

define the anti-collapse regularization term as

$$\mathcal{L}_{ar} = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim S_\mathcal{T}} \left[\frac{1 - \cos(\tilde{\mathbf{s}}_1, \tilde{\mathbf{s}}_2)}{1 - \cos(\mathbf{z}_1, \mathbf{z}_2)} \right]. \tag{7}$$

We can observe that this term amplifies the dissimilarity of the two fake feature vectors when the latent codes generating them are of high similarity. With the case mode collapse more likely occurs being assigned with higher penalty, the generator is forced to mine minor modes in the feature space during training. The discriminator will also handle fake features from the minor modes. Thus, it is expected that more diverse features can be synthesized when applying the generator on novel classes.

With the above two regularization terms, we reach our final training objective as

$$\min_G \max_D \sum_{i=1}^2 L_{GAN_i} + \alpha \sum_{i=1}^2 L_{cr_i} + \beta \frac{1}{L_{ar}}, \tag{8}$$

where α and β are two hyper-parameters. **Algorithm 1** outlines the main training steps of the proposed method.

3.3. Classification with Synthesized Samples

In the test stage, given an FSL task $\mathcal{T}' = (S'_\mathcal{T}, Q'_\mathcal{T})$ randomly sampled from the test set that the classes have no overlap with those in the training set, we first augment the labeled support set $S'_\mathcal{T}$ with the learned generator G. Then, we train a classifier with the augmented

supported set. The classifier is used to classify samples from the query set $Q'_\mathcal{T}$. Specifically, suppose after data augmentation, we get an enlarged support set $\hat{S}'_\mathcal{R} = \{(\mathbf{s}_1, y_1), (\mathbf{s}_2, y_2), \cdots, (\mathbf{s}_{N \times K'}, y_{N \times K'})\}$ where K' is the number of samples synthesized for each class. With $\hat{S}'_\mathcal{R}$, we train a standard Softmax classifier $\mathbf{f}_c$ as

$$\min_\theta \mathop{\mathbb{E}}_{(\mathbf{s},y)\sim \hat{S}'_\mathcal{R}} \log[-P(y|\mathbf{s}; \theta)], \tag{9}$$

where θ is the parameter of $\mathbf{f}_c$. With $\mathbf{f}_c$, we classify samples from $Q'_\mathcal{T}$.

4. Experiments

We evaluate AFHN on three common benchmark datasets, namely, *Mini-ImageNet* [39], *CUB* [40] and *CI-FAR100* [19]. The *Mini-ImageNet* dataset is a subset of ImageNet. It has 60,000 images from 100 classes, 600 images for each class. We follow previous methods and use the splits in [32] for evaluation, i.e., 64, 16, 20 classes as training, validation, and testing sets, respectively. The *CUB* dataset is a fine-grained dataset of totally 11,788 images from 200 categories of birds. We use the split in [17] and 100, 50, 50 classes for training, validation, and testing, respectively. The CIFAR-100 dataset contains 60,000 images from 100 categories. We use the same data split as in [47]. In particular, 64, 16 and 20 classes are used for training, validation and testing, respectively.

Following previous methods, we evaluate 5-way 1-shot and 5-way 5-shot classification tasks where each task instance involves classifying test images from 5 sampled classes with 1 or 5 randomly sampled images for each class as the support set. In order to reduce variance, we repeat the evaluation task 600 times and report the mean of the accuracy with a 95% confidence interval.

4.1. Implementation Details

Following the previous data augmentation based methods [35, 6, 4], we use ResNet18 [16] as our feature extraction network F. We implement the generator G as a two-layer MLP, with LeakyReLU activation for the first layer and ReLU activation for the second one. The dimension of the hidden layer is 1024. The discriminator is also a two-layer MLP, with LeakyReLU as the activation function for the first layer and Sigmoid for the second layer. The dimension of the hidden layer is also 1024. The noise vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ are drawn from a unit Gaussian with the same dimensionality as the feature embeddings.

Following the data augmentation based FSL methods [35, 6], we perform two-step training procedures. In the first step, we only train the feature extraction network F as a multi-class classification task using only the training split. We use Adam optimizer with an initial learning rate 10^{-3}

cWGAN	✗	✗	✓	✓	✓
CR	✗	✓	✗	✓	✓
AR	✗	✗	✗	✗	✓
	52.73	55.65	57.58	60.56	62.38

Table 1. Ablation study on the *Mini-ImageNet* dataset for the 5-way 1-shot setting. cWGAN, CR, and AR represent the conditional WGAN framework, the classification regularizer, and the anti-collapse regularizer, respectively. The baseline result (52.73) is obtained by applying the SVM classifier directly on ResNet18 features without data augmentation. The result (55.65) with only CR added is obtained from the synthesized features produced by the generator without the discriminator and AR during training.

which decays to the half every 10 epochs. We train F with 100 epochs with batch size of 128. In the second training stage, we train the generator and discriminator alternatively, using features extracted by F and update G after every 5 updates of D. We also use Adam optimizer which has an initial learning rate of 10^{-5} and decays to the half every 20 epochs for both F and G. We train the whole network with 100 epochs with 600 randomly sampled FSL tasks in each epoch. For the hyper-parameters, we set $\lambda = 10$ as suggested by [13], and $\alpha = \beta = 1$ for all the three datasets. During the test stage, we synthesize 300 fake features for each class.

The code is developed based on PyTorch.

4.2. Ablation Study

The proposed AFHN consists of the novel conditional WGAN (cWGAN) based feature synthesize framework and the two regularizers that encourage diversity and discriminability of the synthesized features, i.e., the Classification Regularier (CR) and Anti-collapse Regularizer (AR). To evaluate the effectiveness and impact of these components, we conduct ablation study on the *Mini-ImageNet* dataset for the 5-way 1-shot setting. The results are shown in Table 1.

CR. This regularizer constrains the synthesized features to have desirable classification property such that we can train from them a discriminative classifier. We can see that when it is used as the only regularization for the generator, it raises the baseline result from 52.73 to 55.65. On the other hand, when it is used along with cWGAN (the discriminator regularizes the generated features, resulting in the GAN loss), it helps further boost the performance from 57.58 to 60.56. Therefore, in the both cases (with and without cWGAN), CR helps enhance discriminability of the synthesized features and leads to performance boost.

cWGAN. Compared with the baseline (without data augmentation), cWGAN helps raise the accuracy from 52.73 to 57.58. This is because the synthesized features enhance the intra-class variance, which makes classification decision boundaries much sharper. Moreover, with CR as the regularizer, our cWGAN based generative model boosts the

performance of the naive generative model from 55.65 to 60.56. This further substantiates the effectiveness of the proposed cWGAN framework. The performance gain is due to the adversarial game between the generator and the discriminator, which enhances the generator's capability of modeling complex data distribution among training data. The enhanced generator is therefore able to synthesize features of both higher diversity and discriminability.

As mentioned in the related work, one of the major differences of the proposed AFHN from the other feature hallucination based FSL method [41] is that AFHN is an adversarial generative model while [41] uses a naive generative model. This study thus evidences the advantage of AFHN over [41].

AR. AR aims to encourage the diversity of the synthesized features by explicitly penalizing the case where mode collapses more likely occur. Table 1 shows that it further brings about 2% performance gains, thus proving its effectiveness.

4.3. Comparative Results

Mini-Imagenet. *Mini-Imagenet* is the most extensively evaluated dataset. From Table 2 we can observe that AFHN attains the new state-of-the-art, for both the 1-shot and 5-shot setting. Compared with the other four data augmentation based methods, AFHN reaches significant improvements: it beats Δ-encoder [35] by more than 8% for the 5-shot setting and Dual TriNet [6] by more than 3% for the 1-shot setting. Compared with MetaGAN [45] which is also based on GAN, AFHN achieves about 10% improvements for both the 1-shot and 5-shot settings. Besides the significant advantages over the peer data augmentation based methods, AFHN also exhibits remarkable advantages over the other two categories of methods. It beats the best metric learning based method DCEM [8] by about 3.5% for the 1-shot setting. It also performs better than the state-of-the-art meta-learning based algorithms. Compared with the baseline method, "ResNet18+SVM", AFHN reaches about 10% and 5% improvements for the 1-shot and 5-shot settings, respectively. This substantiates the effectiveness of our proposed data augmentation techniques.

CUB. This is a fine-grained bird dataset widely used for fine-grained classification. Recently, it has been employed for few-shot classification evaluation. Thus, relatively less results are reported on this dataset. From Table 3 we can see that AFHN reaches comparable results with both the other two data augmentation based methods Dual TriNet and Δ-encoder. It beats the best metric learning based method SAML [14] by 2.4% for the 5-shot setting, and performs significantly better than the meta-learning based methods. Compared with the baseline, we only have a moderate improvement in the 1-shot setting and reach only a marginal boost for the 5-shot setting. We speculate the reason is that this dataset is relatively small, less than 60 images per class

		Backbone	Reference	1-shot	5-shot
	ResNet18 + SVM (baseline)	ResNet18		52.73±1.44	73.31±0.81
MetricL	Matching Net [39]	Conv-64F	NeurIPS'16	43.56±0.84	55.31±0.73
	PROTO Net [36]	Conv-64F	NeurIPS'17	49.42±0.78	68.20±0.66
	MM-Net [2]	Conv-64F	CVPR'18	53.37±0.48	66.97±0.35
	GNN [11]	Conv-256F	Arxiv'17	50.33±0.36	66.41±0.63
	RELATION NET [37]	Conv-64F	CVPR'18	50.44±0.82	65.32±0.70
	DN4 [23]	Conv-64F	CVPR'19	51.24±0.74	71.02±0.64
	TPN [25]	ResNet8	ICLR'19	55.51±0.86	69.86±0.65
	PARN [43]	Conv-64F	ICCV'19	55.22±0.84	71.55±0.66
	SAML [14]	Conv-64F	ICCV'19	57.69±0.20	73.03±0.16
	DCEM [8]	ResNet18	ICCV'19	58.71±0.62	77.28±0.46
MetaL	MAML [9]	Conv-32F	ICML'17	48.70±1.84	63.11±0.92
	META-LSTM [32]	Conv-32F	ICLR'17	43.44±0.77	60.60±0.71
	SNAIL [27]	ResNet-256F	ICLR'18	55.71±0.99	68.88±0.92
	MACO [17]	Conv-32F	Arxiv'18	41.09±0.32	58.32±0.21
	DFSVL [12]	Conv-64F	CVPR'18	55.95±0.89	73.00±0.68
	META-SGD [24]	Conv-32F	Arxiv'17	50.47±1.87	64.03±0.94
	PPA [31]	WRN-28-10	CVPR'18	59.60±0.41	73.74±0.19
	UFDA [21]	ResNet18	CIKM'19	60.51	77.08
	LEO [34]	WRN-28-10	ICLR'19	61.76±0.08	77.59±0.12
DataAug	MetaGAN [45]	Conv-32F	NeurIPS'18	52.71±0.64	68.63±0.67
	Dual TriNet [4]	ResNet18	TIP'19	58.80±1.37	76.71±0.69
	Δ-encoder [35]	ResNet18	NeurIPS'18	59.90	69.70
	IDeMe-Net [4]	ResNet18	CVPR'19	59.14±0.86	74.63±0.74
	AFHN (Proposed)	ResNet18		**62.38±0.72**	**78.16±0.56**

Table 2. Few-shot classification accuracy on *Mini-Imagenet*. "MetricL", "MetaL" and "DataAug" represent metric learning based category, meta-learning based category and data augmentation based category, respectively. The ± indicates 95% confidence intervals over tasks. The best results are in **bold**.

		Backbone	Reference	CUB		CIFAR100	
				1-shot	5-shot	1-shot	5-shot
	ResNet18 + SVM (baseline)	ResNet18		66.54±0.53	82.38±0.43	59.65±0.78	76.75±0.73
MetricL	Matching Net [39]	Conv-64F	NeurIPS'16	49.34	59.31	50.53±0.87	60.30±0.82
	PROTO Net [36]	Conv-64F	NeurIPS'17	45.27	56.35	-	-
	DN4 [23]	Conv-64F	CVPR'19	53.15±0.84	81.90±0.60	-	-
	SAML [14]	Conv-64F	ICCV'19	69.33±0.22	81.56±0.15	-	-
MetaL	MAML [9]	Conv-32F	ICML'17	38.43	59.15	49.28±0.90	58.30±0.80
	META-LSTM [32]	Conv-32F	ICLR'17	40.43	49.65	-	-
	MACO [17]	Conv-32F	Arxiv'18	60.76	74.96	-	-
	META-SGD [24]	Conv-32F	Arxiv'17	66.90	77.10	61.60	77.90
DataAug	Dual TriNet [6]	ResNet18	TIP'19	69.61	84.10	63.41±0.64	78.43±0.64
	Δ-encoder [35]	ResNet18	NeurIPS'18	69.80±0.46	82.60±0.35	66.70	79.80
	AFHN (Proposed)	ResNet18		**70.53±1.01**	83.95±0.63	**68.32±0.93**	**81.45±0.87**

Table 3. Few-shot classification accuracy on *CUB* and *CIFAR100*. Please refer Table 2 for details.

on average; a large number of classes only have about 30 images. Due to the small scale of this dataset, the intra-class variance is less significant than that of the *Mini-Imagenet* dataset, such that 5 labeled samples are sufficient to capture most of the intra-class variance. Performing data augmentation is less crucial than that for the other datasets.

CIFAR100. This dataset has the identical structure as the *Mini-ImageNet* dataset. Table 3 shows that AFHN performs the best over all the existing methods and the advantages are sometimes significant. AFHN beats Dual TriNet by 5% and 3% for 1-shot and 5-shot respectively. Compared with the best meta-learning based method, we get 7% and 4% improvements for the 1-shot and 5-shot respectively. Compared with the baseline method, AFHN also reach remarkable gains. We reach about 10% and 5% improvements for 1-shot and 5-shot respectively. This great improvement convincingly substantiates the effectiveness of our GAN based data augmentation method for solving the FSL problem.
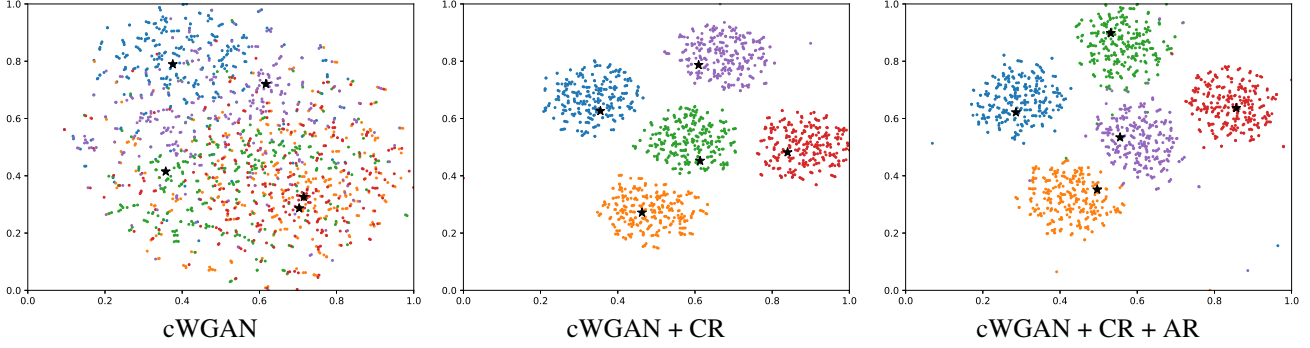
Figure 2. t-SNE [26] visualization of synthesized feature embeddings. The real features are indicated by ⋆. Different colors represent different classes.
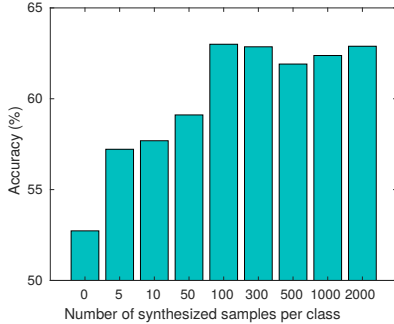

Figure 3. Impact of the number of synthesized samples for each class on the *Mini-ImageNet* dataset.

In summary, among all the three datasets, we reach significant improvements over existing state-of-the-art methods for two of them, while being comparable for the left one. For all the datasets, our method reaches significant boost to the baseline method where there is no data augmentation. These experiments substantiate the effectiveness and superiority of the proposed method.

4.4. Further Analysis

Impact of the number of synthesized features. Figure 3 shows the analysis on *Mini-ImageNet* about the recognition accuracy with respect to the number of synthesized features for each class during test. We can observe that the classification accuracy keeps boosted with more features synthesized at the beginning, and remains stable with even more synthesized samples. This is reasonable because the class variance encapsulated by the few labeled samples has a upper bound; data augmentation based on these labeled samples can enlarge the variance to some extent, but it is still bounded by the few labeled samples themselves. When it reaches the peak, the performance reasonably turns stable.

Visualization of synthesized features. We showed quantitatively in the ablation study that owing to the CR and AR regularizers, we can generate diverse and discriminative features which bring significant performance gains. Here we further study the effect of the two regularizers by show-

ing the t-SNE visualization of the synthesized features. As shown in Figure 2, the synthesized features of different classes mix up together when using only cWGAN for augmentation. As analyzed before, cWGAN does not guarantee synthesizing semantically meaningful features. The problem is substantially resolved when we train cWGAN with CR. The synthesized features exhibit clear clustering structure, which helps train a discriminative classifier. Furthermore, with AR added, the synthesized features still exhibit favorable clustering structure. But taking a closer look of the visualization, we can find that the features synthesized with AR added are more diverse than that without it: the clusterings are less compact, stretched to larger regions, and even contains some noises. This shows AR indeed helps diversify the synthesized features.

5. Conclusions

We introduce the Adversarial Feature Hallucination Networks (AFHN), a new data augmentation based few-shot learning approach. AFHN consists of a novel conditional Wasserstein GAN (cWGAN) based feature synthesis framework, the classification regularizer (CR) and the anti-collapse regularizer (AR). Based on cWGAN, our framework synthesizes fake features for new classes by using the features of the few labeled samples as the conditional context. CR secures feature discriminability by requiring the synthesized features to be of high similarity with features of the samples from the same classes, while of low similarity with those from the different classes. AR aims to enhance the diversity of the synthesized features by directly penalizing the cases where the mode collapse problem likely occurs. The ablation study shows the effectiveness of the cWGAN based feature synthesis framework, as well as the two regularizers. Comparative results verify the superiority of AFHN to the existing data augmentation based FSL approaches as well as other state-of-the-art ones.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

[2] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. Memory matching networks for one-shot image recognition. In *CVPR*, 2018.

[3] Zitian Chen, Yanwei Fu, Kaiyu Chen, and Yu-Gang Jiang. Image block augmentation for one-shot learning. In *AAAI*, 2019.

[4] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. Image deformation meta-networks for one-shot learning. In *CVPR*, 2019.

[5] Zitian Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, Xiangyang Xue, and Leonid Sigal. Semantic feature augmentation in few-shot learning. *arXiv preprint arXiv:1804.05298*, 2018.

[6] Zitian Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, Xiangyang Xue, and Leonid Sigal. Multi-level semantic feature augmentation for one-shot learning. *IEEE Transactions on Image Processing*, 2019.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.

[8] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Diversity with cooperation: Ensemble methods for few-shot classification. In *ICCV*, 2019.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[10] Hang Gao, Zheng Shou, Alireza Zareian, Hanwang Zhang, and Shih-Fu Chang. Low-shot learning via covariance-preserving adversarial augmentation networks. In *NeurIPS*, 2018.

[11] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*, 2017.

[12] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *CVPR*, 2018.

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017.

[14] Fusheng Hao, Fengxiang He, Jun Cheng, Lei Wang, Jianzhong Cao, and Dacheng Tao. Collect and select: Semantic alignment metric learning for few-shot learning. In *ICCV*, 2019.

[15] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *CVPR*, 2017.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[17] Nathan Hilliard, Lawrence Phillips, Scott Howland, Artëm Yankov, Courtney D Corley, and Nathan O Hodas. Few-shot learning with metric-agnostic conditional embeddings. *arXiv preprint arXiv:1802.04376*, 2018.

[18] Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. Edge-labeling graph neural network for few-shot learning. In *CVPR*, 2019.

[19] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[20] Kai Li, Zhengming Ding, Kunpeng Li, Yulun Zhang, and Yun Fu. Support neighbor loss for person re-identification. In *ACM MM*, 2018.

[21] Kai Li, Martin Renqiang Min, Bing Bai, Yun Fu, and Hans Peter Graf. On novel object recognition: A unified framework for discriminability and adaptability. In *CIKM*, 2019.

[22] Kai Li, Martin Renqiang Min, and Yun Fu. Rethinking zero-shot learning: A conditional visual classification perspective. In *ICCV*, 2019.

[23] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In *CVPR*, 2019.

[24] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

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

[26] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[27] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017.

[28] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *ICML*, 2017.

[29] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *CoRR, abs/1803.02999*, 2018.

[30] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, 2018.

[31] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, 2018.

[32] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

[33] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[34] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *ICLR*, 2019.

[35] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Rogerio Feris, Abhishek Kumar, Raja Giryes, and Alex M Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition. *arXiv preprint arXiv:1806.04734*, 2018.

[36] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[37] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018.

[38] Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. Few-shot learning through an information retrieval lens. In *NeurIPS*, 2017.

[39] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, 2016.

[40] Catherine Wah, Steve Branson, Pietro Perona, and Serge Belongie. Multiclass recognition and part localization with humans in the loop. In *ICCV*, 2011.

[41] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *CVPR*, 2018.

[42] Yu-Xiong Wang and Martial Hebert. Learning to learn: Model regression networks for easy small sample learning. In *ECCV*, 2016.

[43] Ziyang Wu, Yuwei Li, Lihua Guo, and Kui Jia. Parn: Position-aware relation networks for few-shot learning. In *ICCV*, 2019.

[44] Aron Yu and Grauman Kristen. Low-shot learning via covariance-preserving adversarial augmentation networks. In *ICCV*, 2017.

[45] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In *NeurIPS*, 2018.

[46] Handong Zhao, Zhengming Ding, and Yun Fu. Multi-view clustering via deep matrix factorization. In *AAAI*, 2017.

[47] Fengwei Zhou, Bin Wu, and Zhenguo Li. Deep meta-learning: Learning to learn in the concept space. *arXiv preprint arXiv:1802.03596*, 2018.