One-Shot Face Recognition via Generative Learning

Zhengming Ding[†], Yandong Guo[♯], Lei Zhang[♯] and Yun Fu[†‡]
[†] Department of Electrical & Computer Engineering, Northeastern University, USA
[♯] Microsoft Research
[‡] College of Computer & Information Science, Northeastern University, USA

***Abstract*— One-shot face recognition measures the ability to recognize persons with only seeing them once, which is a hallmark of human visual intelligence. It is challenging for existing machine learning approaches to mimic this way, since limited data cannot well represent the data variance. To this end, we propose to build a large-scale face recognizer, which is capable to fight off the data imbalance difficulty. To seek a more effective general classifier, we develop a novel generative model attempting to synthesize meaningful data for one-shot classes by adapting the data variances from other normal classes. Specifically, we formulate conditional generative adversarial networks and the general Softmax classifier into a unified framework. Such a two-player minimax optimization can guide the generation of more effective data, which benefit the classifier learning for one-shot classes. The experimental results on a large-scale face benchmark with 21K persons verify the effectiveness of our proposed algorithm in one-shot classification, as our generative model significantly improves the recognition coverage rate from 25.65% to 94.84% at the precision of 99% for the one-shot classes, while still keeps an overall Top-1 accuracy at 99.80% for the normal classes.**

I. Introduction

Face recognition [1], [2], [3], [4] in the wild has been improved with a large margin with the gains from deep convolutional networks (ConvNets) that learn rich feature representations [5], [6], [7]. Modern face recognition models take advantage of the large labeled face datasets, (e.g., MS-Celeb-1M [8], MegaFace [9]) to build good visual representations and train powerful classifiers (*representation learning*). It is now well studied that an effective way to build a face recognition system given thousands of face samples per person is to train a deep ConvNet with Softmax classifier as the final layer [5], [10], [11]. Specifically, Softmax classifier has the following advantages over k-nearest neighbor classifier. One is that the computing complexity of estimating the persons' identity after feature extraction is linear to the number of persons, not the number of images in the gallery. The second one is that the weight vectors for each class are estimated through the information from all the classes. Moreover, Softmax is not very sensitive to data labeling quality in the gallery set. However, Softmax classifier suffers from the data imbalance situations, where some classes would have tons of training samples while other classes have only very few or one training sample (Figure 1).

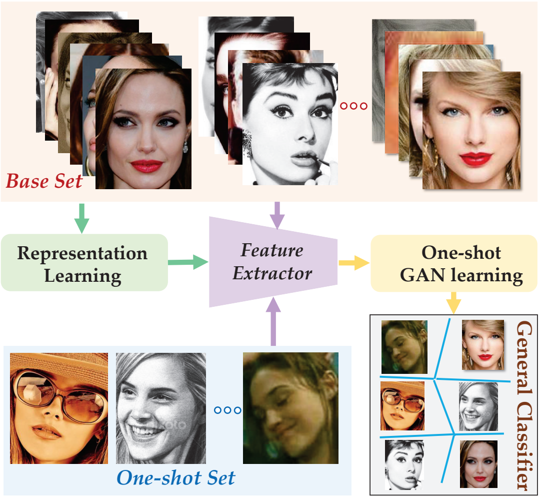


Fig. 1. Illustration of one-shot face recognition problem with two phases. **Representation learning** phase seeks general face visual knowledge through training effective feature extractor using the base set. **One-shot GAN learning** phase builds a general classifier to recognize persons in both the base set and the one-shot set based on the deep features.

In fact, there is a common long-tail phenomenon (serious data imbalance issue) in the existing large-scale face datasets. Moreover, we also confront such difficult situations in real-world applications that we have limited training samples for some specific persons, especially when the number of persons to be recognized is extremely large. From the observations of [12], Softmax classifier outputs very poor performance in recognizing persons in the one-shot set, since the persons in the one-shot set only have one training image per person. The authors notice that one-shot classes can only occupy a much smaller partition in the feature space, and they further reveal that there exists a close relationship between the volume of a class partition in the feature space and the norm of the weight vector for this class in the Softmax classifier model (we also show the phenomenon in the experiments). Therefore, it is essential to enlarge the feature space of one-shot classes to enhance the general classifier learning for the real-world face recognition with data imbalance issue.

One-shot learning has attracted great attentions, which attempts to make progress towards imparting this human ability to modern recognition systems [13], [14], [15], [12], [10]. Generally, there are two lines of strategies to handle

This work is supported in part by the NSF IIS award 1651902, NIJ Graduate Research Fellowship 2016-R2-CX-0013, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Award W911NF-17-1-0367.

*Corresponding author: Yandong Guo. This work was finished when Zhengming Ding was a research intern at Microsoft Research.

978-1-5386-2335-0/18/$31.00 ©2018 IEEE

such a challenge. One line is data augmentation, the other one is classifier adaptation. Along the first line, some human-designed data generation strategies are adopted to synthesize fake data for the low-shot/one-shot classes to boost the classification ability. Edwards et. al proposed handling the one-shot classification task by learning dataset statistics using the amortized inference of a variational auto-encoder [16]. Hariharan et al. designed a method for hallucinating additional examples for the data-starved novel classes [13]. Mehrotra et al. presented an additional generator network based on the Generative Adversarial Networks where the discriminator is the proposed residual pairwise network [15]. For the second line, the idea is to adapt the classifiers for the one-shot classes through multi-layer transformations or some specific loss to boost the classifier space of one-shot classes. Guo et al. proposed a novel supervision loss named as Underrepresented-classes Promotion (UP) loss term, which aligns the norms of the weight vectors of the one-shot classes (a.k.a. underrepresented-classes) to those of the normal classes [12], which directly boosts the one-shot classifier parameters without considering the difference of class variance for different one-shot classes. However, human designed generation rules cannot well learn the data distribution to synthesize the fake data. Therefore, the existing one-shot research efforts fail to jointly seek a general classifier when automatically generating augmented data.

In this paper, we develop a novel and effective face recognition system with an appealing performance for one-shot classes while keeping base classes at a very high Top-1 accuracy. We target at seeking a general classifier for a combination of the *base classes* (many samples per class) and the *one-shot classes* (one sample per class). The one-shot classes are also named as novel classes in [13]. The core idea of our model is to generate effective auxiliary data for the one-shot classes, and thus, we can span the feature space for the one-shot classes to facilitate the one-shot face recognition. To our best knowledge, this is the first work to explore the generative model in a general classifier learning for the one-shot challenge. Specifically, we focus on the one-shot learning stage, where we attempt to train a more powerful classifier for both base and novel classes. The main contributions of our paper are listed in three folds as follows:

- We jointly incorporate generative adversarial networks in training a general classifier for both base and novel classes. In detail, the generator attempts to synthesize more effective fake data for the one-shot classes to enrich the data space of one-shot classes, while the discriminator is built to guide the face data generation to mimic the data variation of base classes and adapt to generate novel classes.
- We design generative adversarial networks in the feature domain, which we first obtain by training a deep ConvNet model on the base classes. More specifically, we build the conditional generative adversarial networks with an auxiliary classifier to augment more effective features and enhance the general classifier learning for one-shot classes. [1]
- We evaluate our proposed model on a large-scale one-shot face dataset [12], and achieve significant improvement in the one-shot classification with coverage rate 94.84% at the precision of 99%. Meanwhile, our model can still achieve very appealing performance as Top1 accuracy of 99.80% for the base classes.

II. Related Work

In this part, we briefly review two categories of related works, one is one-shot learning and the other one is generative models. Meanwhile, we will highlight the differences between our methods and the existing ones.

A. One-shot Learning

Overall, one-shot learning is still an open problem, whose goal is to learning one-shot classes by adapting knowledge obtained from familiar classes. It attempts to mimic the human cognitive to transfer previously-learned knowledge when recognizing novel classes.

One category of approaches to one-shot learning explores generative models of appearance that tap into a global [19] or a super-category level [20] prior. Generative models based on strokes or partshave provided promising results in restricted domains such as hand-written characters [21]. Dixit et al. leveraged a corpus with attribute annotations to generate additional examples by varying attributes [22]. Hariharan et al. also proposed a way to generate additional examples [13], while our model is non-parametric and directly generates feature vectors. Jia et al. presented an appealing alternative to generation using Bayesian reasoning to infer an object category from a few examples [23]. Among discriminative approaches, early work attempted to utilize a single image of the novel class to adapt classifiers from similar base classes [24], [25] through simple hand-crafted features. Bertinetto et al. regressed from single examples to a classifier [26], while Wang and Hebert [14] regressed from classifiers trained on small datasets to classifiers trained on large datasets.

B. Generative Adversarial Networks

Generative adversarial networks (GANs) [27] contains two neural networks trained in opposition to each other. The generator $G(\cdot)$ takes as input a random noise vector z and synthesizes a fake image $G(z)$. The discriminator $D(\cdot)$ receives as input either a real image or a synthesized image from the generator and outputs a probability distribution over possible image sources. The discriminator is trained to maximize the log-likelihood it assigns to the correct source.

The basic GANs framework can be augmented through side information. One strategy is to supply both the generator and discriminator with class labels or latent information to

[1] We train generative model on feature domain instead of image domain for the following reasons. Typically, image synthesis is a more challenging task than image classification/recognition. Especially for the current generative models, it is still an open problem to generate meaningful faces with high quality in many cases [17], [18]. This might be the reason that we don't find existing one-shot learning work using generative models to synthesize face images with very good performance.

obtain conditional samples [17]. Conditional information, e.g., class labels or semantic information, could significantly enhance the quality of synthesized samples [28]. Richer external information such as image captions and bounding box localizations may improve sample quality further [29]. Instead of feeding auxiliary information to the discriminator, one can train the discriminator with reconstructing side information. This is done by modifying the discriminator to contain an auxiliary decoder network that outputs the class label for the training data [30], [31] or a subset of the latent variables from which the samples are synthesized [18]. It is well-known to improve performance on the original task by forcing a model to perform additional tasks [32]. Additionally, an auxiliary decoder could leverage pre-trained discriminators (e.g., image classifiers) for further enhancing the quality of the synthesized images [33]. Motivated by these considerations, we introduce a novel model that combines both strategies for leveraging side information. That is, the model proposed below is class conditional, as well as with an auxiliary classifier to seek an effective and general face recognition system for the real-world applications.

III. Methodology

In this section, we will first introduce the motivation of our proposed model for one-shot face recognition, then present the framework details as well as the training process.

A. Preliminary

The objective of one-shot face recognition is to measure the recognition ability of a model across classes with one training sample. Specifically, the model is trained on labeled training data with two sets without identity overlap, i.e., **Base Set** (i.e., normal classes) $\{X_b, Y_b\}$ with c_b classes and **Novel Set** (i.e., one-shot classes) $\{X_n, Y_n\}$ with c_n classes. The goal is to build a general c-class recognizer ($c = c_b + c_n$). In this paper, we will mainly focus on the performance on the novel classes while keeping an eye on the accuracy for the base classes.

B. Motivation

One-shot face recognition is challenging due to limited samples during model training, and thus, it is essential to generate more effective data to improve the ability of the general classifier. Traditional data augmentation strategies [13] only adopt human designed rules to generate more data for the one-shot classes, so the enhancement to the classifier is limited. Another challenge is that it usually hurts the base classification when we try to improve the classification ability for one-shot classes. That is, the learned classifier is impractical in real-world applications when dealing with a general face recognition problem. Hence, it is essential to balance these two sets.

Moreover, generative models are very popular due to its promising ability to synthesize effective data, which is similar to the real data with a guidance from a discriminator. While for one-shot face recognition, it is essential to generate effective data with large variations for the one-shot classes in order to span their classifier space. Generally, data from the base classes have large intra-class variations, and thus, it is helpful to adapt the variations of base classes to the one-shot classes during data generation. To generate more meaningful data for one-shot classes, we jointly seek a general classifier with the input of real data and fake data. Such a joint learning framework could benefit generating meaningful data and improving the classification ability.

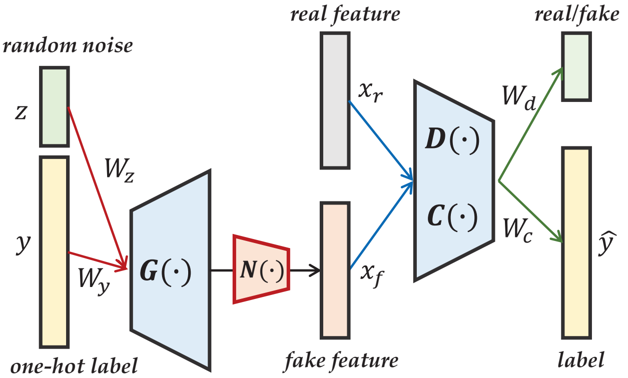


Fig. 2. Illustration of generative one-shot face recognizer, where z is the random noise vector, y is the one-hot label, x_r is the real feature, while x_f is the generated fake feature. $G(\cdot)$ is the generator with the input of random noise z and one-hot label y. The output of generator with normalization $N(\cdot)$ will achieve the fake feature x_f. $D(\cdot)$ is the discriminator which aims to differentiate the real and fake features, while $C(\cdot)$ is the a general multi-class classifier.

C. Generative One-Shot Learning

The goal of generative models manages to synthesize fake data to augment the training data from random noise [27]. Conditional generative models prove to be more effective to synthesize meaningful data [17]. We denote random noise $z \in \mathbb{R}^{d_z}$, and real feature $x \in \mathbb{R}^{d_x}$ with its one-hot label $y \in \mathbb{R}^{d_y}$.

To the generator, the prior input are the combination of random noise $p_z(z)$ and one-hot label y. While to the discriminator, x and y are the input to a discriminative function. Thus, a two-player minimax game can be formulated as follows:

$$\mathcal{L}_d^f = \mathbb{E}[\log(1 - D(G(z|y)))] \tag{1}$$

$$\mathcal{L}_d^r = \mathbb{E}[\log(D(x))] \tag{2}$$

where the generator aims to make the generated features similar to real features, attempting to minimize $\mathcal{L}_d^f$; the discriminator aims to differentiate the real and fake features by maximizing $\mathcal{L}_d^r + \mathcal{L}_d^f$.

Since we mainly focus on the one-shot learning stage, and thus we adopt the discriminative features extracted at the representation learning stage as the input. Suppose the extracted features $X_r \in \mathbb{R}^{d \times m}$, which contains both base classes c_b and one-shot classes c_n, d is the feature dimension and m is the training sample size.

Specifically, the generator attempts to synthesize fake feature x_f with the input of random noise $z \in \mathbb{R}^{d_z}$ and one-hot label $y \in \mathbb{R}^c$. Therefore, we design $G(\cdot)$ as:

$$\begin{aligned} G(z|y) &= f_1(W_g \begin{bmatrix} z \\ y \end{bmatrix}) = f_1([W_z, W_y] \begin{bmatrix} z \\ y \end{bmatrix}) \\ &= f_1(W_z z + W_y y), \end{aligned} \quad (3)$$

where $W_g = [W_z, W_y]$ while $W_z \in \mathbb{R}^{d \times d_z}$ and $W_y \in \mathbb{R}^{d \times c}$. And $f_1(\cdot)$ is the element-wise activation function, e.g., ReLU function or Sigmoid function.

The discriminator manages to differentiate the fake features and real features $D(\cdot)$, which is designed as

$$D(x) = f_2(W_d x), \quad (4)$$

where $W_d \in \mathbb{R}^{1 \times d}$ and $f_2(\cdot)$ projects x to a scale between 0 and 1.

Our goal is to seek a general classifier $C(\cdot)$ with c classes as:

$$\mathcal{L}_c^r = \mathbb{E}[\log P(Y = y|x)] \quad (5)$$

where we target at seeking a general classifier weight matrix $W_c \in \mathbb{R}^{c \times d}$.

However, for the c_n one-shot classes, there are limited samples for training a good classifier. Motivated by GANs [27], we hope the generated samples can be used to boost the classifier learning. Thus, we have the classifier loss for fake features as:

$$\mathcal{L}_c^f = \mathbb{E}[\log P(Y = y|G(z|y))]. \quad (6)$$

As shown above, the discriminator $D(\cdot)$ is trained through maximizing $\mathcal{L}_c^r + \mathcal{L}_d^f + \mathcal{L}_d^r$, while the generator $G(\cdot)$ is trained through maximizing $\mathcal{L}_c^f - \mathcal{L}_d^f$. Thus, $D(\cdot)$ and $G(\cdot)$ both manage to maximize $\mathcal{L}_c^r + \mathcal{L}_c^f$.

Remark: For the generator $G(\cdot)$, we attempt to synthesize meaningful data to augment the one-shot classes. The goal is to span the feature space of one-shot classes around its center features. Generally, we can calculate the averaged feature of base classes as their centers, while the one-shot features for novel classes are usually not the centers and may be far away from their centers. Thus, we hope $W_y y$ could keep the class center information, while $W_z z$ to capture the class variation information. Therefore, we initialize W_y with the class-center features, and then we would get its class center by multiplying W_y and its one-hot label y. Specifically, the base part is initialized with the averaged features of c_b classes, while novel part is initialized with the available one-shot features. For the random noise part, we aim to mimic the class variations. Thus, we initialize W_z and z randomly. In this way, we can capture the data variation within base classes and adapt to generate more meaningful data for novel classes. Note that the scale of $W_y y$ is the same as that of x, and thus, the scale would be improved if we add a random part $W_z z$. Therefore, we add a normalization process to make the fake feature to have the same scale as that of the the real feature. Specifically, $N(x_f) = x_f / \|x_f\|_2 * \alpha$, where α is the averaged norm of real feature.

For $C(\cdot)$, W_c are the classifier parameters for both the base and novel classes. We initialize the classifier parameters trained on the base and novel dataset with the ResNet-34 deep features [34] (See detail in experiments). As known to all, a deep model training on base classes with many samples per class can achieve very promising results for base classes [12]. That is, the classifier parameters are good enough for base classes recognition. The goal of one-shot learning is to improve the classifier parameters for one-shot classes. Hence, we hope the base classifier parameters of W_c to be similar to the pre-trained one. We develop a square loss regularizer to constrain the base classifier not far away from its initialized one. In this way, not only can we update the classifier parameters for novel classes to enhance the classification ability, but also relax the classifier space for base classes, triggering the expansion of novel classifier space.

We implement our model with TensorFlow[2]. We set the learning rate as 10^{-4} and the optimizer as Adam optimizer. We adopt leaky-relu and sigmoid activation functions for $G(\cdot)$ and $D(\cdot)$, respectively. Since GANs can be solved as a *minimax* optimization problem, we first constrain the generator to optimize the discriminator, then fix the discriminator to update the generator. Thus, we alternatively update two neural networks until the model converges.

IV. Experimental Results

In this section, we first introduce the one-shot face data as well as its feature representation process. Then we provide the one-shot evaluations with other comparisons to verify the effectiveness of our proposed model. Finally, we go deep and show some phenomena of our model.

A. One-Shot Face Dataset

The face dataset[3] used here is sampled from MS-Celeb-1M dataset [8]. In total, this dataset contains 21K people with 1.2M images, which is considerably larger than other publicly available datasets except for the MS-Celeb-1M dataset. To evaluate the one-shot challenge, we divide the dataset into base set (20K) parts, i.e., base set (20K people) and novel set (1K people). Since we want to build a general 21K-class classifier for both the base and novel classes, we hope our optimized classifier achieve promising performance on both sets, otherwise it is meaningless in the real-world applications.

In the base set, there are 20K persons, each of which having 50-100 images for training and 5 for test. In the novel set, there are 1000 persons, each with one image for training and 10 for test. The experimental results in this paper were obtained with 100K test images for the base set and 20K test images for the novel set. We focus on the recognition performance in the novel set, while monitoring the recognition performance in the base set to ensure that the performance improvement in the novel set does not harm the performance in the base set.

To recognize the test images for the persons in the novel set is a challenging task. The one training image per

[2] https://www.tensorflow.org/

[3] http://www.msceleb.org/challenge2/2017

TABLE I

Coverage at Precisions = 99% and 99.9% on the one-shot set, where our generative model significantly improves the coverage at precision 99% and 99.9%.

Method	C@P=99%	C@P=99.9%
Fixed-Feature	25.65%	0.89%
SGM [13]	27.23%	4.24%
Update Feature	26.09%	0.97%
Direct Train	15.25%	0.84%
Shrink Norm[12]	32.58%	2.11%
Equal Norm[12]	32.56%	5.18%
Up Term [12]	77.48%	47.53%
Ours	94.84%	83.82%

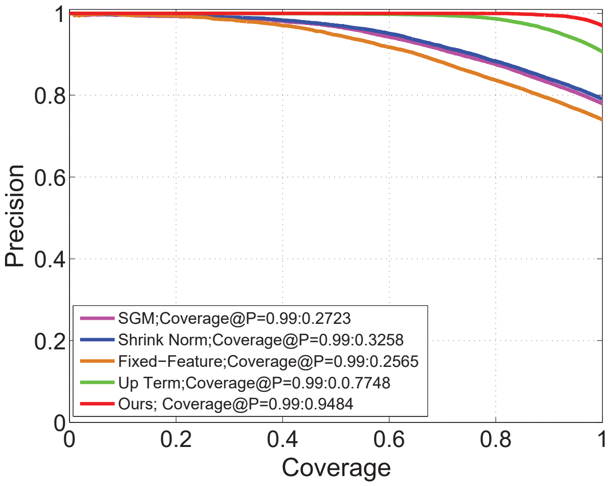


Fig. 3. Precision-Coverage curves of five methods on the one-shot set, where our model achieves a very appealing coverage@precision=99%.

person was randomly preselected, and the selected image set includes images of low resolution, profile faces, and faces with occlusions. The training images in the novel set show a large range of variations in gender, race, ethnicity, age, camera quality (or evening drawings), lighting, focus, pose, expressions, and many other parameters.

Representation learning: we train deep ResNet-34 model with input faces' resolution as 224×224 and seek a 20K-class classifier using all the training images of the 20K persons in the base set. There are about 50-100 images per person in the base set. The wrong labels in the base set are very limited (less than 1% based on manual check). We crop and align face areas to generate the training data[4]. Our face representation model is learned from predicting the 20K classes. More specifically, we consider each person as one class and train a deep convolutional neural network (ConvNet) supervised by the Softmax with the cross-entropy loss. We have tried different network structures and adopted the standard residual network with 34 layers [34] due to its good trade-off between prediction accuracy and model complexity. Features extracted from the last pooling layer are adopted as the face representation (512 dimensions).

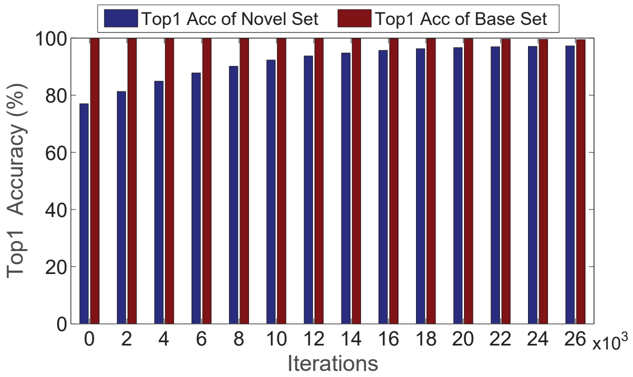


Fig. 4. Top1 accuracy (%) of base set and novel set with different iterations, where we notice that our model could significantly improve the Top1 accuracy for the novel classes while keeping a very promising Top1 accuracy for the base classes.

B. *One-shot Face Recognition*

We compare with the following algorithms:

- **Fixed-Feature**: updates the feature extractor and only train the Softmax classifier with the feature extractor provided by phase one.
- **Updated Feature**: fine-tunes the feature extractor simultaneously when we train the Softmax classifier in phase two. The feature updating does not change the recognizer's performance too much.
- **SGM** [13]: is known as squared gradient magnitude loss, is obtained by updating the feature extractor during phase one using the feature shrinking method.
- **Shrink norm** [12]: adopts L_2-norm to shrink classifier parameters, which is one typical strategy to handle insufficient data problem efficiency.
- **Equal norm** [12]: is a weight regularizer, which constrains the classifier parameters of both novel and base classes to the same value.
- **UP Term** [12]: is a weight regularizer, which only enforces the classifier parameters of the novel classes to the same value.

All the methods are based on a 21K-class classifier (trained with different methods). Note that we boost all the samples in the novel set for 100 times for all the methods, since the largest number of samples per person in the base set is about 100.

The experimental results of our method and the alternative methods are listed in Table I. We adopt coverage rate at precision 99% and 99.9% as our evaluation metrics since this is the major requirement for a real recognizer [12].

Compared with the Fixed-Feature, SGM method obtains around 2% improvements in recall when precision is 99%, while 4% improvements when precision requirement is 99.9%. The gain for face recognition by feature shrinking in [13] is not as significant as that for general image. The reason might be that the face feature is already a good representation for faces and the representation learning is not a major bottleneck. Note that we did not apply the feature hallucinating method as proposed in [13] for fair comparison

[4] http://www.msceleb.org/download/lowshot

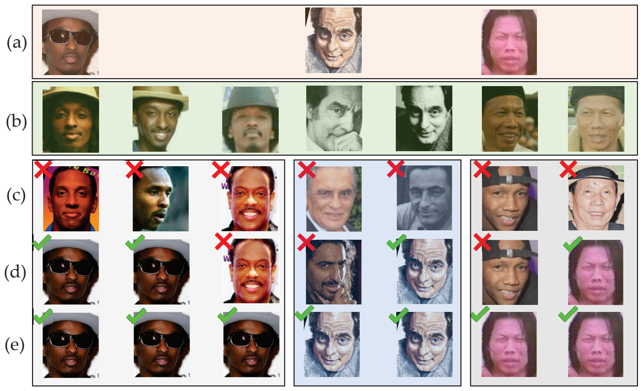


Fig. 5. Face retrieval results, where row (a) denotes the three challenges one-shot training faces, i.e., occlusion, sketch, low-resolution. Row (b) represents the test images, while the the bottom three rows show the recognized results of three models, i.e., (c) KNN, (d) Softmax, and (e) Our generative model.

and to highlight the contribution of model learning, rather than data augmentation. To couple the feature hallucinating method (may need to be modified for face) is a good direction for the next step.

Our model significantly improves the one-shot classification, preserving base classification at a very promising performance. Specifically, as shown in Table I, our generative model improves the coverage@precision=99% and coverage@precision=99.9% significantly. Moreover, we notice that our model can achieve the state-of-the-art performance without any external data by comparing the competitors in the low-shot challenge [5]. This verifies our generative model is able to synthesize very effective features to alleviate the one-shot classification.

The coverage at precision 99% on the base set obtained by using any classifier-based methods in Table I is 100%. The Top1 accuracy on the base set obtained by any of these classifier-based methods is $99.80 \pm 0.02\%$. Thus, we do not report them separately in the table. That verifies that our generative model could synthesize meaningful one-shot samples to boost classifier space for one-shot classes.

C. Face Retrieval Results

We select three typical one-shot training cases (Figure 5), i.e., low-solution, sketch, occlusion, to quantitatively show the performance of different models. We compare with k-nearest neighbor classifier (KNN) ($k = 1$), Softmax, and our model. All the models are input with the pre-trained deep ResNet-34 features.

From the results (Figure 5), we observe that our model can well handle these three challenging cases and recognize these persons correctly. KNN cannot correctly recognize the testing images of these three persons, which results from that the testing images are quite different from the one-shot training image. Softmax can retrieve some correct ones, which shows more promising results than that of KNN. Hence, we consider KNN is not suitable in one-shot face classification in large-scale dataset. Our model can significantly handle those three challenging cases, which results from the generation of effective data in facilitating the classifier learning.

D. Property Analysis

First of all, we evaluate the Top1 accuracy of the base set and novel set with the model optimization. From the results 4, we observe that the Top1 accuracy of one-shot set is significantly improved from 77.01% to 96.24%. This shows that our model enhances the classification for one-shot classes by spanning the feature space. We further notice that the classification accuracy for the base set is hurt somehow, but very slightly. That demonstrates our generative model can learn a good general classifier, which is much practical in real-world scenarios.

Secondly, we present more information for the classifier to deeply understand why our model can improve the one-shot classification. Specifically, we have a c-class classifier, with each weight vector w in W_c for each class. Thus, we evaluate the norm of each class weight vector to see the variations of these information. From the results (Figure 6), we notice that from (a) to (f) with more iterations' optimization, the norms of the weight vectors corresponding to the novel classes are triggered to similar distribution as that of the base classes (f). Actually, (a) shows the results of the initialized parameters obtained from Softmax trained on ResNet-34 deep features. That is the reason we consider why our model significantly improves the one-shot classification, since we boost the variance of novel classes to be similar to base classes. Such phenomenon is also obtained in [12], where the authors states that the norm of classifier weight is related to the classifier space.

V. Conclusions

In this paper, we proposed a generative framework for one-shot face recognition, where we attempted to synthesize more effective augmented data for one-shot classes by borrowing the data variation of base set. Specifically, generative learning was jointly incorporated in the general classifier training for both the base and one-shot classes. Thus, more effective fake data were generated for the one-shot classes to enrich the data space of one-shot classes. Furthermore, a discriminator was designed to guide the face data generation to mimic the data variation of base classes and adapt to generate one-shot classes. Experiments on a large-scale one-shot face benchmark showed that our model could significantly improve the performance of one-shot classification, while keeping the promising classification ability for the base set.

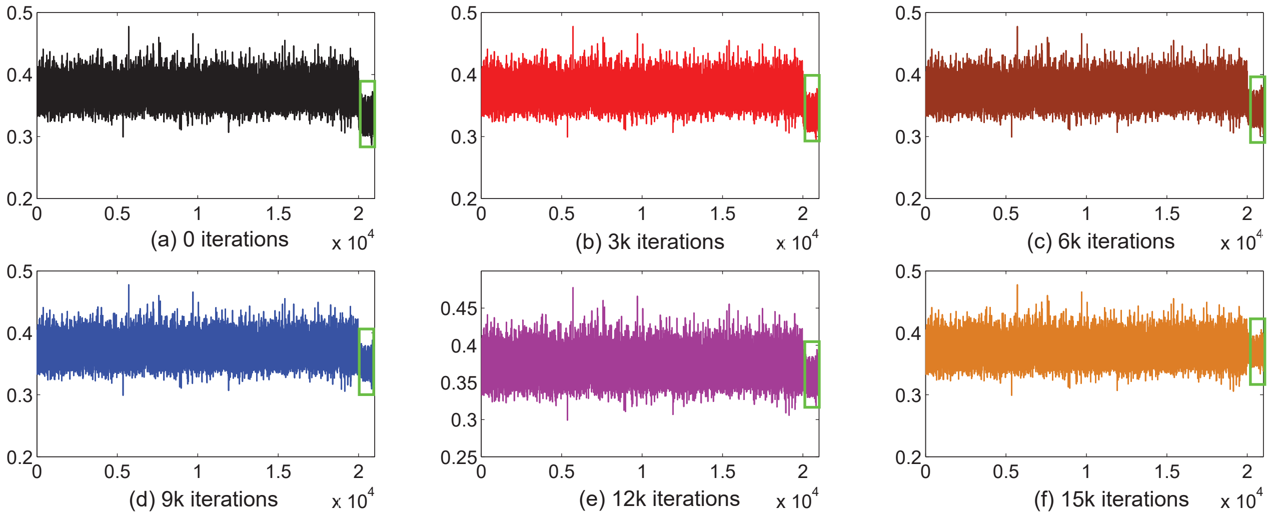
References

[1] H. Zhao, Z. Ding, and Y. Fu, "Block-wise constrained sparse graph for face image representation," in *IEEE International Conference on Automatic Face and Gesture Recognition*, vol. 1. IEEE, 2015, pp. 1–6.

[2] Z. Ding, S. Suh, J.-J. Han, C. Choi, and Y. Fu, "Discriminative low-rank metric learning for face recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*, vol. 1. IEEE, 2015, pp. 1–6.

[3] Z. Ding, M. Shao, and Y. Fu, "Latent low-rank transfer subspace learning for missing modality recognition," in *Association for the Advancement of Artificial Intelligence*, 2014, pp. 1192–1198.

[5] http://www.msceleb.org/leaderboard/c2



(a) 0 iterations

(b) 3k iterations

(c) 6k iterations

(d) 9k iterations

(e) 12k iterations

(f) 15k iterations

Fig. 6. Norm of the classifier weight vector w for each class in W_c. The x-axis is the class index. The rightmost 1000 classes on the x-axis correspond to the persons in the novel set. As shown in the figure, with more iterations from (a) to (f), $\|w\|_2$ for the novel set tends to have similar values as that of the base set (Green bounding box denotes the weights for one-shot classes). This promotion introduces significant performance improvement.

[4] S. Wang, Z. Ding, and Y. Fu, "Coupled marginalized auto-encoders for cross-domain multi-view learning." in *International Joint Conference on Artificial Intelligences*, 2016, pp. 2125–2131.

[5] Y. Wu, J. Li, Y. Kong, and Y. Fu, "Deep convolutional neural network with independent softmax for large scale face recognition," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1063–1067.

[6] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[7] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *12th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 2017, pp. 118–126.

[8] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 87–102.

[9] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4873–4882.

[10] Y. Wu, H. Liu, and Y. Fu, "Low-shot face recognition with hybrid classifiers," in *The IEEE International Conference on Computer Vision (ICCV) Workshop*, Oct 2017.

[11] Y. Xu, Y. Cheng, J. Zhao, Z. Wang, L. Xiong, K. Jayashree, H. Tamura, T. Kagaya, S. Shen, S. Pranata, J. Feng, and J. Xing, "High performance large scale face recognition with multi-cognition softmax and feature retrieval," in *The IEEE International Conference on Computer Vision (ICCV) Workshop*, Oct 2017.

[12] Y. Guo and L. Zhang, "One-shot face recognition by promoting underrepresented classes," *arXiv preprint arXiv:1707.05574*, 2017.

[13] B. Hariharan and R. Girshick, "Low-shot visual object recognition," *arXiv preprint arXiv:1606.02819*, 2016.

[14] Y.-X. Wang and M. Hebert, "Learning from small sample sets by combining unsupervised meta-training with cnns," in *Advances in Neural Information Processing Systems*, 2016, pp. 244–252.

[15] A. Mehrotra and A. Dukkipati, "Generative adversarial residual pairwise networks for one shot learning," *arXiv preprint arXiv:1703.08033*, 2017.

[16] H. Edwards and A. Storkey, "Towards a neural statistician," *arXiv preprint arXiv:1606.02185*, 2016.

[17] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[18] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2016, pp. 2172–2180.

[19] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[20] R. Salakhutdinov, J. Tenenbaum, and A. Torralba, "One-shot learning with a hierarchical nonparametric bayesian model," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 195–206.

[21] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[22] M. Dixit, R. Kwitt, M. Niethammer, and N. Vasconcelos, "Aga: Attribute guided augmentation," *arXiv preprint arXiv:1612.02559*, 2016.

[23] Y. Jia and T. Darrell, "Latent task adaptation with large-scale hierarchies," in *IEEE International Conference on Computer Vision*, 2013, pp. 2080–2087.

[24] E. Bart and S. Ullman, "Cross-generalization: Learning novel classes from a single example by feature replacement," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2005, pp. 672–679.

[25] A. Opelt, A. Pinz, and A. Zisserman, "Incremental learning of object detectors using a visual shape alphabet," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2006, pp. 3–10.

[26] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *Advances in Neural Information Processing Systems*, 2016, pp. 523–531.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[28] A. v. d. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with pixelcnn decoders," *arXiv preprint arXiv:1606.05328*, 2016.

[29] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.

[30] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.

[31] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[32] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[33] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 3387–3395.

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