Demodalizing Face Recognition with Synthetic Samples

Zhonghua Zhai[1,2*] **Pengju Yang**[1*] **Xiaofeng Zhang**[1] **Maji Huang**[1]
Haijing Cheng[1] **Xuejun Yan**[1] **Chunmao Wang**[1] **Shiliang Pu**[1†]
[1]Hikvision Research Institute
[2]Zhejiang University
{zhaizhonghua, yangpengju, zhangxiaofeng15, huangmaji, chenghaijing,
yanxuejun, wangchunmao, pushiliang.hri}@hikvision.com

Abstract

Using data generated by generative adversarial networks or three-dimensional (3D) technology for face recognition training is a theoretically reasonable solution to the problems of unbalanced data distributions and data scarcity. However, due to the modal difference between synthetic data and real data, the direct use of data for training often leads to a decrease in the recognition performance, and the effect of synthetic data on recognition remains ambiguous. In this paper, after observing in experiments that modality information has a fixed form, we propose a demodalizing face recognition training architecture for the first time and provide a feasible method for recognition training using synthetic samples. Specifically, three different demodalizing training methods, from implicit to explicit, are proposed. These methods gradually reveal a generated modality that is difficult to quantify or describe. By removing the modalities of the synthetic data, the performance degradation is greatly alleviated. We validate the effectiveness of our approach on various benchmarks of large-scale face recognition and outperform the previous methods, especially in the low FAR range.

Introduction

Deep face recognition aims to map an input image to a feature space with a small intraclass distance and a large interclass distance. Previous work was implemented via loss function design and datasets with rich intraclass differences (Schroff, Kalenichenko, and Philbin 2015; Wen et al. 2016; Liu et al. 2017; Wang et al. 2018; Deng et al. 2019). However, most of the face datasets that we have utilized are biased and long-tailed. Even very large public datasets manifest strong biases in image characteristics, such as ethnicity (Sohn et al. 2018), age (Wen, Li, and Qiao 2016; Zheng, Deng, and Hu 2017) and head poses (Masi et al. 2016; Peng et al. 2017). On the other hand, a few classes tend to have rich in-class samples, while most classes have very few in-class samples (Yin et al. 2019; Liu et al. 2019). These biased and long-tailed characteristics greatly affect the performance of the recognition model, especially on difficult test datasets.

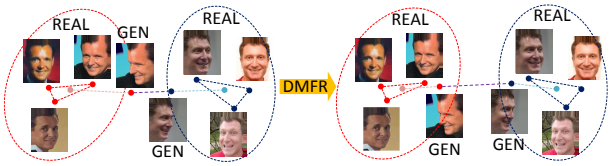

Figure 1: Due to changes in attributes (such as the head pose and image quality), the distance between synthetic samples and real samples in the same category may be relatively large, and due to the existence of a fixed synthetic mode, the distance between the synthetic samples of different people will be close. After DMFR removes the synthetic modality, the synthetic samples will be closer to the real samples in their own category and farther from the generated samples in other categories. This figure is best viewed in color.

With continuous improvement in the quality of synthetic images, some researchers have begun to use generative adversarial networks (GANs) (Goodfellow et al. 2014) and three-dimensional (3D) methods (Feng et al. 2018; Masi et al. 2016) to generate images to enrich classes with fewer samples. This approach is applied in the fields of person re-identification (re-ID) (Dai et al. 2018; Zheng et al. 2019) and few-shot learning (Mishra et al. 2018; Mondal, Dolz, and Desrosiers 2018). However, in face recognition tasks, we discover that directly mixing real and synthetic samples in training often yields negative effects (Shi et al. 2020) regardless of how realistic the synthetic samples appear. After experimentation, we discovered that the synthetic samples often have unified modality information. As with adversarial samples, this information often does not have obvious visual characteristics, but it will create problems for training the model.

For the first time, we propose demodalizing face recognition (DMFR) to remove fixed modalities in synthetic samples. As shown in Fig. 1, the distribution of the synthetic samples after removing the modal information is similar to that of the real samples, and the distances between the synthetic samples increases. The classifier does not need to make incorrect offsets to fit sample points that are outliers due to the modal information. Therefore, the synthetic samples can be better utilized for face recognition and other tasks.

*Equally-contributed
†Corresponding author

Specifically, we propose three learning methods, from implicit to explicit. The first learning method is the meta-learning-based method. Our model implicitly learns methods for removing modal information from the generation domain and can learn to generalize well to unseen attacks. The second learning method is the disentangling-based method. We disentangle the feature that contains only identity information from the other information, which contains the modality, and then use the identity information feature for identification. This process eliminates the influence of the modal information of the synthetic sample on the recognition training. The third method is the filter-based method. We propose a filter structure to explicitly remove the modalities in the synthetic samples so that the synthetic features can filter out the modal information after passing through the filter structure without losing the other information. In order to highlight our theme, we select three plain methods as backbones and remove the more fancy techniques, such as cross reconstruction and information loss in disentangle-based method.

In **Experiments** Section, we discuss how synthetic modalities affect the recognition model training on the basis of feature differences and similarity distribution changes. We provide the insightful analysis that removing the synthetic modal information will improve the learning of the identity embeddings. Comprehensive experiments demonstrate that the use of face recognition benchmarks, such as Labeled Faces in the Wild(LFW)(Huang et al. 2008), YouTube Faces(YTF)(Wolf, Hassner, and Maoz 2011), CFP-FP (Sengupta et al. 2016) and MegaFace (Kemelmacher-Shlizerman et al. 2016), substantially improve our model performance compared to the direct addition of synthetic samples. In some challenging test datasets, such as the IARPA Janus Benchmark-A(IJB-A) (Klare et al. 2015), IARPA Janus Benchmark-C(IJB-C) (Maze et al. 2018), CP-IJB-C and CQ-IJB-C, a new state-of-the-art performance is achieved. After synthetic samples remove modal information, the samples produce effects in the corresponding domain (such as a large posture).

The main contributions of this paper are presented as follows:

- For the first time, we highlight the synthetic modal problem, which has a great influence on real samples.
- We propose a novel demodalizing face recognition (DMFR) framework to solve the synthetic modal problem, which removes the modal information of synthetic samples during training.
- We extract samples from IJB-C and build a cross-pose dataset CP-IJB-C and a cross-quality dataset CQ-IJB-C.
- We achieve state-of-the-art results on several challenging benchmarks, such as IJB-A, IJB-C, CP-IJB-C and CQ-IJB-C.

Related Work

Deep Face Recognition

Deep neural networks have been extensively applied in face recognition research. FaceNet (Schroff, Kalenichenko, and Philbin 2015) proposes triplet loss to maximize the distance between the anchor and negative samples while minimizing the distance between the anchor and positive samples. Center loss (Wen et al. 2016) aims to minimize the distances between samples and their class centers. Marginal loss (Deng, Zhou, and Zafeiriou 2017) introduces the concept of the margin, which uses the margin value to limit the distances between classes while minimizing the distances within classes. SphereFace (Liu et al. 2017) proposed angular softmax loss (A-Softmax). CosFace (Wang et al. 2018) adopted an additive cosine margin, and Arcface (Deng et al. 2019) adopted an additive angular margin.

Data Augmentation

Deep learning models often rely heavily on data. Therefore, data augmentation is extensively applied to increase the amount of training data. Data augmentation generally includes flipping, rotating, and resizing. In addition to these general data augmentation methods, 3D generative models (Feng et al. 2018; Masi et al. 2016) and GANs (Goodfellow et al. 2014) are employed in fields such as person re-ID (Dai et al. 2018; Zheng et al. 2019) and few-shot learning (Mishra et al. 2018; Mondal, Dolz, and Desrosiers 2018).

Meta Learning

The meta-learning method focuses on 1) learning a better initialization weight to quickly adapt to new tasks, such as model-agnostic meta-learning (MAML) (Finn, Abbeel, and Levine 2017) and its variant Reptile (Nichol, Achiam, and Schulman 2018), as well as meta transfer learning (Sun et al. 2019) and iMAML (Rajeswaran et al. 2019); 2) learning an embedding space and a classifier, which can be employed to directly classify samples in a new task without fast adaptation (Vinyals et al. 2016; Snell, Swersky, and Zemel 2017; Sung et al. 2018); and 3) after pretraining a feature extractor on an entire training set, learning the parameters of the predictive classifier (Qiao et al. 2018; Gidaris and Komodakis 2018). Most of these works focus on few-shot learning with a small number of categories, and the application of meta-learning is very challenging for face recognition tasks that involve thousands of people.

Disentangled Representations

Disentangling representation aims to learn how to separate identity information from irrelevant information. In a real scenario, the learned model is applied to extract identity information as the identification features. Early feature disentangling is mostly based on artificially designed features (Gong et al. 2013). With the development of neural networks, features extracted by convolutional neural networks have begun to replace manually designed features, different loss functions were designed, and end-to-end feature disentangling was realized. With the development of GAN, generation-based methods (Tran, Yin, and Liu 2017) and feature disentangling methods begin to combine.

Modality-invariant Feature Learning

Modality-invariant feature learning aims to extract the modality-invariant features. Since modal changes are re-

moved when extracting or learning these features, these features are only related to facial identity information and are quite robust to changes in modality. Traditional methods of modality-invariant feature learning are generally based on manually extracted features. Among deep learning methods, Wu et al. pro-posed a cross-modal ranking named coupled deep learning (CDL) (Wu et al. 2017), which reduces the domain discrepancy. Huo et al. proposed a discriminative feature learning method (Huo et al. 2017). He et al. employed the Wasserstein distance to reduce the gap between domains to obtain domain-invariant features (He et al. 2018).

Proposed Approach

Modality of the Synthetic Data

In the deep learning era, each sample x_i is represented as an embedding z_i in a latent space, that is, $z_i = H(x_i)$. The conditional synthetic sample feature has the form

$$x_i' = x_i \oplus \Delta x_i = x_i \oplus g(x_i, \Delta c) \oplus \epsilon(x_i)$$

where g represents the generation function, Δc represents the difference in the generation conditions, such as the pose angle, age or clarity, and we assume that

$$\Delta c = 0 \Leftrightarrow g(x_i, \Delta c) = 0$$

$\epsilon(x_i)$ represents the fixed mode of the synthetic samples, which we refer to as the modality of the synthetic samples, and $\oplus$ represents a nonlinear relationship. Here, we assume the simplest case that the synthetic modality is additive. In the ideal generation situation, we hope that $\epsilon(x_i) = 0$, that is, that the difference between the synthetic sample and the real sample is related only to the term $g(x_i, \Delta c)$. With this ideal assumption, when $\Delta c = 0, x_i' = x_i$. In **Experiments** Section, we show that the synthetic modality $\epsilon(x_i)$ exists. This modality explains why it is difficult for us to directly use the synthetic samples for training.

Synthetic Sample Demodalization

We propose three demodalization methods, from implicit to explicit. Meta-learning is a modality-invariant method. When it is applied to the training of synthetic samples, we treat the real samples and synthetic samples as different domains and meta-learning as a non-explicit demodalization method. In the disentangling method, we aim to disentangle the ID information and non-ID information of the face and ultimately use the ID information for face recognition. After disentangling, demodalization is performed on the synthetic sample, even if the modality is still coupled in the non-ID information. In the filter method, we propose an end-to-end method to directly remove the modal information from the synthetic samples. Unlike the previous two methods, the demodalization in this method is not a black box but a module that can be controlled manually. We can also obtain explicit modality information by filters.

Meta-learning-based DMFR

Overview In the training phase, we have several domains, the real domain D_{tr} and N generative domains $D_{g_1}, D_{g_2}, \ldots, D_{g_N}$. Each domain contains a certain face label. In the testing phase, the trained model is evaluated on several domains $D_{T_1}, D_{T_2}, \ldots, D_{T_M}$, which may not have been previously observed. In addition, the labels of the test domain and training domain are disjoint, so we are addressing an open set problem. We propose a meta-learning-based DMFR to use the generalization ability of meta-learning to implicitly eliminate the modal information in the generative domains so that the model can achieve better performance in the test domain.

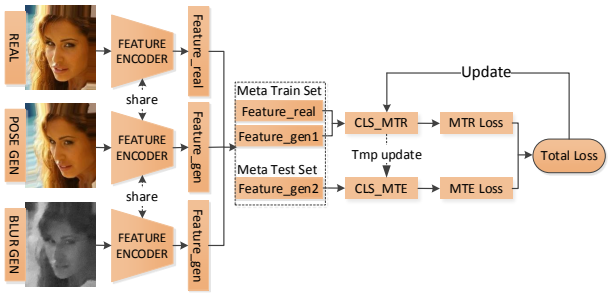

Figure 2: Overview of the proposed meta-learning-based DMFR model. The figure is best viewed in color.

Domain-based Sampling To remove the modal information of the synthetic samples, we divide the training domain into two parts in each training iteration: meta-training and meta-testing. Specifically, we randomly select a generative domain for the meta-test and use the remainder of the generative domain and real domain D_{tr} for the meta-training. Therefore, the real domain and generative domain shift can be simulated. Unlike the MAML (Finn, Abbeel, and Levine 2017) method, to remove the modal information of the generative domain, we add the real domain to the meta-training in each training iteration instead of completely randomly selecting the meta-training and meta-testing sets. In this way, our model can autonomously learn a method for removing the modal information from the generative domain and can learn to generalize well to unseen attacks.

Meta-optimization A convolutional neural network that is composed of a feature extractor and a meta-learner is proposed in meta-learning-based DMFR. We investigate the posterior of the probability being classified to identity $j \in \{1, 2, ..., I\}$, given the input sample x_i. Denote the feature embedding of sample i as $\mathbf{f}_i$ and j_{th} identity prototype vector as $\mathbf{w}_j$. The whole meta-optimization procedure is illustrated in Fig. 2; the details are presented as follows:

Based on domain-based sampling, during each batch within a meta-batch, we sample N-1 source domains, which are composed of one real domain and N-2 generative domains. We calculate the classification loss in each batch as

$$\mathcal{L}_{tr} = \frac{1}{I}\sum_{1}^{I} -\log \frac{\exp s(\mathbf{w}_{y_i}^T \mathbf{f}_i - m)}{\exp s(\mathbf{w}_{y_i}^T \mathbf{f}_i - m) + \sum_{j \neq y_i} \exp s(\mathbf{w}_j^T \mathbf{f}_i)}$$

where y_i is the ground-truth label of x_i. The prototype $\mathbf{w}$ is next updated by gradients $\nabla_{\mathbf{w}}$ as:

$$\mathbf{w}' = \mathbf{w} - \alpha \nabla_{\mathbf{w}} \mathcal{L}_{tr}$$

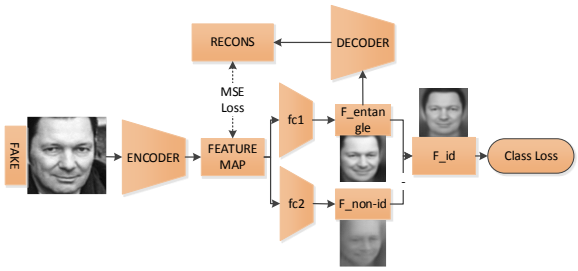

Figure 3: Overview of the proposed disentangling-based DMFR model. The figure is best viewed in color.

Moreover, we sample batches in the remaining generative domain for the meta-test. We encourage our model trained on the meta-train domain to perform well on the meta-test domain by removing the modal information of the generative domains. The loss is then calculated on the updated parameters $\mathbf{w}'$ as follows:

$$\mathcal{L}_{te} = \frac{1}{I}\sum_{1}^{I} -\log \frac{\exp s(\mathbf{w'}_{y_i}^T \mathbf{f}_i - m)}{\exp s(\mathbf{w'}_{y_i}^T \mathbf{f}_i - m) + \sum_{j \neq y_i} \exp s(\mathbf{w'}_j^T \mathbf{f}_i)}$$

To simultaneously optimize the meta-training and meta-testing, the final objective is

$$argmin_\theta \lambda_{meta}\mathcal{L}_{tr}(\theta, \mathbf{w}) + (1-\lambda_{meta})\mathcal{L}_{te}(\theta, \mathbf{w} - \alpha\nabla_{\mathbf{w}}\mathcal{L}_{tr})$$

where θ is the parameter of the feature extractor, α is the learning rate of meta-train and λ_{meta} is the hyperparameter that balances the meta-training and meta-testing. In each round of training, a gradient is back-propagated on the meta-training, while a meta-gradient is back-propagated on the meta-testing. Since the meta-training domain always contains real samples, the modal information of the generative domain in the process of gradient back-propagation is implicitly filtered so that the model performs well in both the meta-training domain and the meta-testing domain.

Disentangle-based DMFR

Overview The goal of disentangling is to split features that contain identity information only from other information that includes modalities and use the identity information features for face recognition. For the synthetic samples $x_i' = x_i \oplus g(x_i, \Delta c) \oplus \epsilon(x_i)$, we propose a disentangling method that can eliminate the influence of the modal information on the recognition training. We divide the entangled feature $f_{entangle}$ into two parts, f_{id} and f_{non-id}, where f_{id} is the identity information used for identification, $f_{non-id} = H(g(x_i, \Delta c) \oplus \epsilon(x_i))$, H is the feature extractor and $\epsilon(x_i)$ is the modal information that needs to be removed. The remaining issue is that the two parts $g(x_i, \Delta c)$ and $\epsilon(x_i)$ are still coupled. The whole disentangling procedure is illustrated in Fig. 3. Specifically, we use the classification loss to constrain f_{id} only containing the information needed to predict the class identity. In our experiment, we additionally train a decoder that visualize the features, and we find that f_{id} abandon non-identity information, such as posture, to achieve better recognition results. In this situation, to recover a reconstructed image with the same posture as the input sample (we use a reconstruction loss to ensure this process), $f_{entangle}$ has to contain rich information, such as posture. Through a minus operation, we compress non-ID information into f_{non-id}, which can be clearly seen from the visualization image in Fig. 3.

Feature Map Reconstruction Disentangling-based DMFR has an encoder-decoder structure. The encoder structure is divided into two parts. The first part extracts a relatively informative feature $f_{entangle}$, followed by a decoder to ensure that $f_{entangle}$ is capable of recovering the original feature map. Note that our encoder-decoder structure does not recover the original input data but is designed to recover a feature map $\mathbf{M}$ of appropriate depth. First, lower-level neural network operations require more computing resources, and second, a feature map of appropriate depth already contains most of the information that we need and is able to complete the work of the encoder-decoder structure. Blindly seeking to restore the original image would introduce too much averaging information because of the reconstruction loss function. The loss function of reconstruction is

$$\mathcal{L}_{recons} = \|\mathbf{M}_{recons} - \mathbf{M}\|_2$$

where $\mathbf{M}_{recons}$ is the reconstruction of the selected feature map.

The other part of the encoder structure extracts the feature f_{non-id}, which is unrelated to the target classification task; it contains the synthetic modal information with which we are most concerned, although this part of the information is still coupled with other information, such as the pose, illumination and age. The remainder of our networks are designed and trained to ensure that f_{id} contains only identity information and is thus separated from f_{non-id}, which contains only domain and other information.

Classification Disentangling-based DMFR contains a classifier to constrain f_{id} to be maximally informative about the identity information while eliminating the maximum amount of other irrelevant information, that is, to maximize the removal of synthetic modal information. We calculate the classification loss in each batch as follows:

$$\mathcal{L}_{cls} = \frac{1}{I}\sum_{1}^{I} -\log \frac{\exp s(\mathbf{w}_{y_i}^T \mathbf{f}_i - m)}{\exp s(\mathbf{w}_{y_i}^T \mathbf{f}_i - m) + \sum_{j \neq y_i} \exp s(\mathbf{w}_j^T \mathbf{f}_i)}$$

Optimization Disentangling-based DMFR is an end-to-end framework, and the final objective is a linear combination of all loss functions:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{disen}\mathcal{L}_{recons}$$

where λ_{disen} is a hyperparameter that is chosen experimentally.

Filter-based DMFR

Filter We propose the filter structure F to explicitly remove the modality from the synthetic samples. The goal is that the synthetic features will filter out the modal information after passing through the filter structure without losing other information. To achieve this goal, we generated a batch

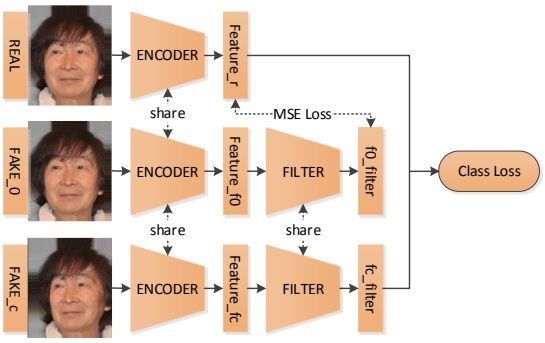

Figure 4: Overview of the proposed filter-based DMFR model. The figure is best viewed in color.

of samples with $\Delta c = 0$, where c represents the generation conditions, such as the angle, age and clarity. When $\Delta c = 0$, the difference between the synthetic samples and the real samples should originate from only the synthetic modality.

Specifically, we define the representation in the latent space of each sample x_i as $H(x_i; \theta_h)$, where θ_h represents the parameters of the feature extraction network H. The pairwise feature distance between $H(x_i; \theta_h)$ and $F(H(x_i \oplus g(x_i, \Delta c) \oplus \epsilon(x_i); \theta_h))$ is characterized by the Euclidean distance and can be formulated as follows:

$$\mathcal{L}_{flt} = \|F(H(x_i \oplus g(x_i, \Delta c) \oplus \epsilon(x_i); \theta_h)) - H(x_i; \theta_h)\|_2$$

Classification Loss Synthetic features will be filtered by Filter F and then fed into a classifier with the real features to minimize the following classification loss:

$$\mathcal{L}_{cls} = \frac{1}{I}\sum_{1}^{I} - \log \frac{\exp s(\mathbf{w}_{y_i}^T \mathbf{f}_i - m)}{\exp s(\mathbf{w}_{y_i}^T \mathbf{f}_i - m) + \sum_{j \neq y_i} \exp s(\mathbf{w}_j^T \mathbf{f}_i)}$$

In practice, we use different classifiers for real data and synthetic data. The experimental results show that this experimental setting is helpful for improving the recognition performance.

Optimization Last, we use

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{flt}\mathcal{L}_{flt}$$

as the total loss function, where λ_{flt} is a trade-off hyperparameter.

Experiments

Datasets and Implementation Details

In this section, we describe the public datasets that we employed and provide some implementation details.

Training Datasets We use a cleaned version of the MS-Celeb-1M datasets (Guo et al. 2016) with 2,251,420 images of 58,982 subjects as our training set. Note that we follow the lists (Wang et al. 2019a,b) to remove the overlapped identities between the employed training datasets and the test datasets.

Synthetic Datasets Our model for generating samples of different poses is based on PRNet (Feng et al. 2018), which according to the two-dimensional planar structure of a Basel face model (BFM) (Paysan et al. 2009) parameterized on the plane, directly selects a UV map that contains 53,215 points and establishes a new 3D-to-2D mapping matrix. The modeling accuracy exceeds that of the original PRNet.

To evaluate the influence of synthetic data generated by other methods, we use cycle-GAN (Zhu et al. 2017) to generate blurred synthetic samples and use them in meta-learning based and disentangling-based DMFR along with the synthetic samples generated by PRNet.

Architecture We construct a face image(137×169) by warping a face region using three facial points: the two eyes and the midpoint of the two corners of the mouth. We employ the modified 100-layer ResNet (He et al. 2016) as the backbone network.

In disentangling-based DMFR, the encoder produces feature maps with the spatial size 13×13 and a depth of 1024 channels. The feature maps are then divided in depth into two trunks, which are dedicated to rich information and non-ID information. The two trunks are then passed to fully connected layers to generate the final representations $f_{entangle}$ and f_{non-ID}, with both cardinalities set to $d = 512$. The decoder is a deconvolution network.

Training We train the model with 8 synchronized graphic processing units (GPUs) and a mini-batch, including 128 images per GPU. To make the training more stable, all DMFR networks are based on a network that is pretrained by only softmax loss. We use an initial learning rate of 0.01 and reduce the learning rate by 0.1 at 50k, 70k and 80k with a weight decay of 0.0005 and a momentum of 0.9 using stochastic gradient descent (SGD).

Hyperparameters We empirically set $\lambda_{meta} = 0.5$, $\lambda_{disen} = 1.0$ and $\lambda_{flt} = 1.0$, respectively. The margin m is empirically set to 0.4.

Analysis

Synthesis Sample Modal Visualization To verify that a synthetic sample has a fixed mode, we use the optimized PRNet (Feng et al. 2018) to generate artificial samples, whose attitude angle changes to Δc, where $c \in \{0, 0.1, 1, 10\}$, as shown in Figure 5. The real sample and the synthetic samples are visualized at the image and feature levels. When Δc is very small, it is difficult to observe changes that are visible to the naked eye at the image level. However, from the visualized image of the feature, we observe that in some feature dimensions, there is a natural discriminability between the real sample and the synthetic sample, which is the modality of the generated samples.

Similarity of Synthesis Samples We extract the features from the last layer of the baseline model and compute the cosine similarity as the similarity metric. Table 1 presents the results of these experiments. We observe that after adding synthetic samples to the training set, the intraclass similarity decreases. This is reasonable because we increase the richness of the intraclass information by introducing changes in

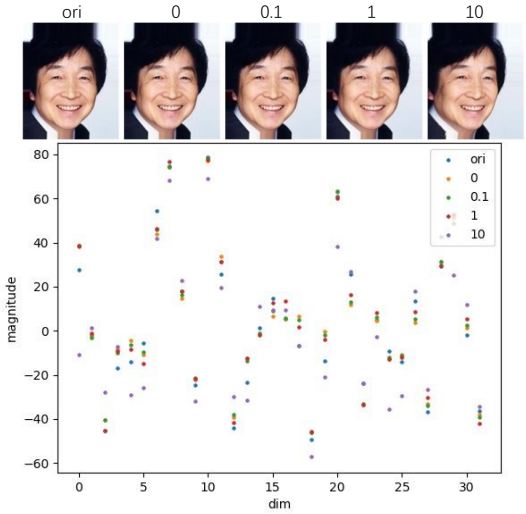

Figure 5: When $\Delta c < 1$, it is difficult to see the difference in the picture, but in terms of the features, there are obvious differences between the original features and the synthesized features in some dimensions. The figure is best viewed in color.

Similarity	Intraclass	Interclass
Baseline	0.809	0.022
Add synthetic samples	0.788	0.216
After filtering	0.782	0.028

Table 1: Fixed modality leads directly to the higher similarity between classes. After the filter, there is no significant change in the similarity within the feature class while the interclass similarity is obviously reduced.

the face pose or image quality. On the other hand, the similarity between classes increases because of the similar fixed modality between the generated samples. This fixed modality leads directly to the model's poor performance after the synthetic samples are added. We also analyze the changes in the similarity of the features after the filter. There is no significant change in the similarity within the feature class after the filter. This proves that the filter retains the richness of the generated samples. On the other hand, the interclass similarity is obviously reduced because the filter has a role in filtering the generated modal information.

Loss Values Analysis We observe that the loss of training with the generated data in a mini-batch at a ratio of $1:1$ (orange curve) decreases more slowly than that of training with the real data (blue curve) and suffers from oscillation. Training with the generated data added to the mini-batch in a ratio of $1:2$ (green curve) cannot converge normally. The loss curves of our proposed method decrease steadily, which indicates the convergence of each DMFR net, as shown in Fig. 6.

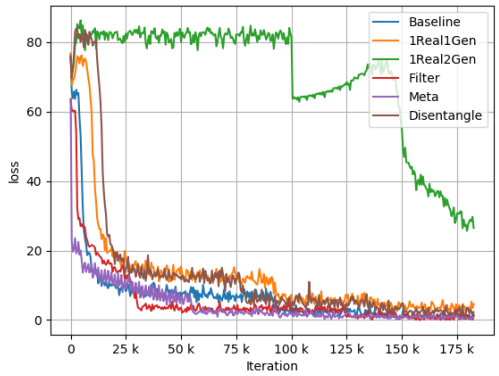

Figure 6: The loss of training with the generated data in a mini-batch at a ratio of $1:1$ (orange curve) decreases more slowly than that of training with the real data (blue curve) and suffers from oscillation. Training with the generated data added to the mini-batch in a ratio of $1:2$ (green curve) cannot converge normally. The loss curves of DMFR methods decrease steadily, indicating the convergence of each DMFR net. The figure is best viewed in color.

Method	LFW	YTF	CFP-FP	MegaFace
w/o synthetic samples	0.9975	0.9490	0.9623	0.7893
CosFace	0.9968	0.9578	0.9737	0.7563
ArcFace	0.9925	0.9650	0.9589	0.6392
DUL(Chang et al. 2020)	0.9973	0.9622	0.9713	0.7804
Shi et al.	0.9962	0.9620	0.9721	0.7810
DMFR(meta)	0.9947	0.9588	0.9640	0.7525
DMFR(filter)	0.9972	0.9542	0.9699	0.7792
DMFR(disentangle)	0.9960	0.9616	0.9633	0.7660

Table 2: In LFW, YTF, CFP-FP and MegaFace, there is no obvious difference between the performance of directly adding the synthetic samples into training and the universal representation or DMFR methods. Only the model of the first row is training without synthetic samples.

Evaluations

Compared Methods The original CosFace (Wang et al. 2018) is employed as the baseline. The classification loss counterparts include CosFace and ArcFace (Deng et al. 2019). In addition, we compare some recent universal representation methods, such as confidence-aware identification loss (Shi et al. 2020) and DUL (Chang et al. 2020), as universal representation counterparts. We reimplement these methods following every detail in their original literature and conduct a fair comparison with the same experimental settings.

Evaluation on General Datasets We compared our method with the baseline and the performance after directly adding generated data on the general face recognition test set, that is, the test set with limited intraclass changes and high quality. Table 2 summarizes the results for these evaluations. Because the quality of most of the test images is satisfactory and our method is mainly applied to large poses and low-quality situations, our method has no advantage in

Method	IJB-A(TAR@FAR)			IJB-C(TAR@FAR)			
	FAR=1e-4	FAR=1e-3	FAR=1e-2	FAR=1e-5	FAR=1e-4	FAR=1e-3	FAR=1e-2
w/o synthetic samples	0.61718	0.85133	0.95418	0.55689	0.76622	0.88966	0.95357
CosFace	0.49875	0.88105	0.96231	0.00803	0.25316	0.80283	0.96344
ArcFace	0.56647	0.76712	0.89080	0.53546	0.72710	0.86511	0.94396
DUL(Chang et al. 2020)	0.65190	0.88329	**0.96758**	0.06984	0.40860	0.82456	0.95378
Shi et al.	0.76512	**0.92686**	0.96661	0.48443	0.76822	0.92453	**0.97009**
DMFR(meta)	0.78969	0.91923	0.95402	0.42921	0.80478	0.92795	0.96385
DMFR(filter)	0.65124	0.86516	0.95434	0.64453	0.81659	0.91001	0.95597
DMFR(disentangle)	**0.86662**	0.91275	0.94297	**0.84819**	**0.90852**	**0.94396**	0.96569

Table 3: The proposed DMFR models achieve consistently better results, especially at a low false acceptance rate, than the other methods when evaluating on challenging datasets IJB-A and IJB-C. When FAR=1e-5 in IJB-C, the performance of many methods(CosFace, DUL) crashes directly. Only the model of the first row is training without synthetic samples.

Method	CQ-IJB-C(TAR@FAR)				CP-IJB-C(TAR@FAR)			
	FAR=1e-5	FAR=1e-4	FAR=1e-3	FAR=1e-2	FAR=1e-5	FAR=1e-4	FAR=1e-3	FAR=1e-2
w/o synthetic samples	0.26264	0.4221	0.60965	0.76148	0.10385	0.31806	0.56268	0.76309
CosFace	0.00033	0.00215	0.28013	0.81419	0.00051	0.00161	0.23115	0.81272
ArcFace	0.212	0.39756	0.60327	0.77532	0.19592	0.39463	0.60647	0.76749
DUL(Chang et al. 2020)	0.02660	0.30234	0.64827	0.84643	0.02118	0.25089	0.6456	0.83237
Shi et al.	0.14744	0.34038	0.65863	0.84693	0.02922	0.27164	0.65416	**0.83297**
DMFR(meta)	0.31954	0.42309	0.6844	0.81966	0.157767	0.27901	0.67932	0.80468
DMFR(filter)	0.37303	0.5266	0.65805	0.77358	0.17101	0.37938	0.62697	0.78198
DMFR(disentangle)	**0.59166**	**0.67777**	**0.77341**	**0.84767**	**0.5393**	**0.65975**	**0.74699**	0.81924

Table 4: Meta-learning-based DMFR implicitly removes the modal information and loses its competitiveness on the more difficult test set. Filter-based and disentangle-based DMFR are methods explicitly removing modal information, which performance are much better than other methods. Only the model of the first row is training without synthetic samples.

these test sets. This finding confirms that there is no obvious domain gap between a test set and a training set of this type. Even without adding generated samples or additional strategies, direct training can achieve excellent performance.

Evaluation on Mixed-Quality Datasets When evaluating challenging datasets, which have a large domain gap with high-quality training datasets, the model that directly adds a generated sample in training and other general state-of-the-art models undergo severe performance degradation. The results in Table 3 indicate that the proposed DMFR models achieve consistently better results, especially at a low false acceptance rate, than the other methods. Comparing to methods of directly using generated samples for training, methods trying to use synthetic samples better, such as confidence-aware identification loss (Shi et al. 2020) and our method, are more competitive.

The CP-IJB-C and CQ-IJB-C Datasets To further verify the effectiveness of the synthetic data and the demodalization method, we classify the samples in IJB-C according to the posture information and construct a new test set CP-IJB-C[1]. We make some modifications to the original test protocol. Only the front image is retained in the gallery set, and only the profile image is retained in the probe set. This test protocol can fully verify the robustness of the model to pose changes. Similar to CP-IJB-C, we construct a test set CQ-IJB-C based on the image quality. This test dataset can fully verify the robustness of the model to image quality changes.

Since meta-learning-based DMFR implicitly removes the modal information, it loses its competitiveness on the more difficult test set, although it still performs better than DUL and Confidence-aware Loss. Since filter-based and disentangle-based DMFR are methods for explicitly removing modal information, their performance on CQ-IJB-C and CP-IJB-C are substantially better than other methods, as shown in Table 4.

Conclusion

In this work, we propose three general learning methods of demodalizing face recognition (DMFR) with synthetic samples; they are based on meta-learning, disentangling and filtering. These three methods provide ways for removing fixed modal information from synthetic samples and use different perspectives, from implicit to explicit. Comprehensive experiments demonstrate that our methods perform better than the compared methods on the most challenging benchmarks. We extract samples from IJB-C and build the cross-pose dataset CP-IJB-C and cross-quality dataset CQ-IJB-C.

References

Chang, J.; Lan, Z.; Cheng, C.; and Wei, Y. 2020. Data Uncertainty Learning in Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5710–5719.

[1]https://github.com/lingjiantian/IJBC-attribute

Dai, P.; Ji, R.; Wang, H.; Wu, Q.; and Huang, Y. 2018. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, volume 1, 2.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Deng, J.; Zhou, Y.; and Zafeiriou, S. 2017. Marginal loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 60–68.

Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; and Zhou, X. 2018. Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network. In *European Conference on Computer Vision*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400* .

Gidaris, S.; and Komodakis, N. 2018. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4367–4375.

Gong, D.; Li, Z.; Lin, D.; Liu, J.; and Tang, X. 2013. Hidden factor analysis for age invariant face recognition. In *Proceedings of the IEEE international conference on computer vision*, 2872–2879.

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

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, 87–102. Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, R.; Wu, X.; Sun, Z.; and Tan, T. 2018. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE transactions on pattern analysis and machine intelligence* 41(7): 1761–1773.

Huang, G. B.; Mattar, M.; Berg, T.; and Learned-Miller, E. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments.

Huo, J.; Gao, Y.; Shi, Y.; Yang, W.; and Yin, H. 2017. Heterogeneous face recognition by margin-based cross-modality metric learning. *IEEE transactions on cybernetics* 48(6): 1814–1826.

Kemelmacher-Shlizerman, I.; Seitz, S. M.; Miller, D.; and Brossard, E. 2016. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4873–4882.

Klare, B. F.; Klein, B.; Taborsky, E.; Blanton, A.; Cheney, J.; Allen, K.; Grother, P.; Mah, A.; and Jain, A. K. 2015. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1931–1939.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 212–220.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2537–2546.

Masi, I.; an Trãn, A. T.; Hassner, T.; Leksut, J. T.; and Medioni, G. 2016. Do we really need to collect millions of faces for effective face recognition? In *European conference on computer vision*, 579–596. Springer.

Maze, B.; Adams, J.; Duncan, J. A.; Kalka, N.; Miller, T.; Otto, C.; Jain, A. K.; Niggel, W. T.; Anderson, J.; Cheney, J.; et al. 2018. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, 158–165. IEEE.

Mishra, A.; Verma, V. K.; Reddy, M. S. K.; Arulkumar, S.; Rai, P.; and Mittal, A. 2018. A generative approach to zero-shot and few-shot action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 372–380. IEEE.

Mondal, A. K.; Dolz, J.; and Desrosiers, C. 2018. Few-shot 3d multi-modal medical image segmentation using generative adversarial learning. *arXiv preprint arXiv:1810.12241* .

Nichol, A.; Achiam, J.; and Schulman, J. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999* .

Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; and Vetter, T. 2009. A 3D face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, 296–301. Ieee.

Peng, X.; Yu, X.; Sohn, K.; Metaxas, D. N.; and Chandraker, M. 2017. Reconstruction-based disentanglement for pose-invariant face recognition. In *Proceedings of the IEEE international conference on computer vision*, 1623–1632.

Qiao, S.; Liu, C.; Shen, W.; and Yuille, A. L. 2018. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7229–7238.

Rajeswaran, A.; Finn, C.; Kakade, S. M.; and Levine, S. 2019. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, 113–124.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Sengupta, S.; Chen, J.-C.; Castillo, C.; Patel, V. M.; Chellappa, R.; and Jacobs, D. W. 2016. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–9. IEEE.

Shi, Y.; Yu, X.; Sohn, K.; Chandraker, M.; and Jain, A. K. 2020. Towards Universal Representation Learning for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6817–6826.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, 4077–4087.

Sohn, K.; Shang, W.; Yu, X.; and Chandraker, M. 2018. Unsupervised domain adaptation for distance metric learning. In *International Conference on Learning Representations*.

Sun, Q.; Liu, Y.; Chua, T.-S.; and Schiele, B. 2019. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 403–412.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1199–1208.

Tran, L.; Yin, X.; and Liu, X. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1415–1424.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*, 3630–3638.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5265–5274.

Wang, X.; Wang, S.; Wang, J.; Shi, H.; and Mei, T. 2019a. Co-mining: Deep face recognition with noisy labels. In *Proceedings of the IEEE international conference on computer vision*, 9358–9367.

Wang, X.; Zhang, S.; Wang, S.; Fu, T.; Shi, H.; and Mei, T. 2019b. Mis-classified vector guided softmax loss for face recognition. *arXiv preprint arXiv:1912.00833* .

Wen, Y.; Li, Z.; and Qiao, Y. 2016. Latent factor guided convolutional neural networks for age-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4893–4901.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, 499–515. Springer.

Wolf, L.; Hassner, T.; and Maoz, I. 2011. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, 529–534. IEEE.

Wu, X.; Song, L.; He, R.; and Tan, T. 2017. Coupled deep learning for heterogeneous face recognition. *arXiv preprint arXiv:1704.02450* .

Yin, X.; Yu, X.; Sohn, K.; Liu, X.; and Chandraker, M. 2019. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5704–5713.

Zheng, T.; Deng, W.; and Hu, J. 2017. Age estimation guided convolutional neural network for age-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1–9.

Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; and Kautz, J. 2019. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2138–2147.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.