Single Sample Face Recognition Based on Identity-Attribute Disentanglement and Adversarial Feature Augmentation

Liang Yao, Fan Liu(✉)*, Zhiquan Ou, Fei Wang, and Delong Chen

College of Computer and Information, Hohai University,
Nanjing 211100, China
{liangyao,fanliu,zhiquanou,fei_wang,chendelong}@hhu.edu.cn

Abstract. To address the issue of facial variation interference, this paper proposes a novel approach for single sample face recognition. Inspired by human visual perception, we introduce an attribute disentanglement module to separate identity features from attribute features using canonical correlation analysis. Due to the lack of attribute labels in the single sample set, we utilize the attribute features of the generic set to construct the SOM attribute space. Then, we fine-tune the network by reducing the distance between the attribute features of single sample and the attribute space. Finally, we use feature adversarial augmentation module to generate more intra-class features and train more robust classifier. Experimental results on AR, LFW and FERET datasets show significant improvements in accuracy and generalization performance compared to other methods.

Keywords: Single-sample Face Recognition, Attribute Disentanglement, Adversarial Feature Generation

1 Introduction

Face recognition is a widely used biometric recognition method due to its non-invasive nature, high accuracy, and convenient data acquisition. However, in real-world applications, face recognition systems often face the challenge of having access to only one sample per identity, which significantly limits their performance. Under such single-sample per person (SSPP) [14] constraint, facial attributes like expression, mustache, hair style, and eyeglasses are strongly coupled with facial identity, posing a significant challenge for learning accurate face representation.

* This work was partially supported by National Nature Science Foundation of China(62372155), Joint Fund of Ministry of Education for Equipment Pre-research(8091B022123), Research Fund from Science and Technology on Underwater Vehicle Technology Laboratory(2021JCJQ-SYSJJ-LB06905), Key Laboratory of Information System Requirements, No:LHZZ 2021-M04, Water Science and Technology Project of Jiangsu Province under grant No.2021063, Qinglan Project of Jiangsu Province.

This challenge leads to the need for performing *identity-attribute disentanglement* during the face representation learning process. In this paper, we aim to make face features more robust to attribute variations, allowing identity information to be captured more precisely. Specifically, based on Canonical Correlation Analysis (CCA), we set up an additional attribute classifier and designed a novel loss function, which aims to lower the correlation coefficients of attribute information and identity information to encourage the separation of these two components.

However, this method only works well on datasets with human labeling of facial attributes, which we utilized as a generic set for representation pretraining. When it comes to single sample sets for downstream recognition, attribute annotations are usually inaccessible. This problem poses a significant challenge for learning identity-attribute disentanglement on single sample sets. To solve this issue, this paper constructs a Self Organizing Map (SOM) [15]-based attribute space using the attribute feature set obtained from the disentanglement result of the generic dataset. Then, we fine-tune the network by reducing the distance between the attribute features of single sample and the attribute space.

Additionally, to further enhance the learning of disentanglement on single sample sets, we propose using adversarial feature augmentation to generate virtual face features. We set up a Generative Adversarial Network (GAN) [17] to learn the probabilistic distribution of facial attribute and identity features, then sample a large number of virtual features to enlarge the training set. With sufficient training instances, learning a decent identity-attribute disentanglement becomes easier.

We evaluated our method on the AR [18], LFW [8], and FERET [19] datasets, achieving high face recognition accuracies of 95.2%, 98.34%, and 99.30%, respectively. This represents an absolute improvement of +0.53%, +0.43%, and +5.40% compared to previous methods. Rigorous ablation experiments prove that both identity-attribute disentanglement and adversarial feature augmentation make noticeable contributions to the overall model performance.

2 Related Work

Existing deep single-sample face recognition methods can be divided into two categories: virtual sample methods and generic learning methods[10–12]. When training deep models directly using a single-sample training set, limited training samples often lead to model over fitting. Therefore, the direct solution is to generate multiple virtual samples or features based on a single training sample to expand the training set, thereby transforming the single-sample problem into a general face recognition problem. Generic learning methods introduce an additional generic sample set with rich intra-class variation information to learn variation information as prior knowledge for the network, improving the accuracy of single-sample face recognition problems.

The key to virtual sample methods is to increase intra-class variation within the samples. Because the newly generated virtual samples are highly correlated

with the original single-sample dataset, their contribution to classification improvement is limited. In recent years, various deep learning-based virtual sample methods have been proposed to better simulate the real intra-class distribution of facial images. Some methods employ novel network architectures and generation processes, using GANs to create virtual samples. For example, Zakharov et al. [1] represent intra-class variations using extracted facial landmarks and employ meta-learning strategies to generate high-quality virtual samples during adversarial training. Tran et al. [2] introduced a Disentangled Representation learning-Generative Adversarial Network (DR-GAN) to separate pose features from the features, enabling pose-controllable face generation.

Some researchers argue that generating virtual images also requires input feature extractors, when remapped to the feature space, may lead to identity information loss. Therefore, several methods based on virtual features have been proposed. For instance, Yin et al. [3] assume that intra-class variations of feature vectors follow a Gaussian distribution, allowing the sampling of different virtual features for individual samples from the corresponding distribution. Min et al.[4] learn intra-class variation information from a general sample set through feature clustering.

The aforementioned methods often treat the features extracted from individual samples as the centers of their respective classes, without considering the inherent variation information carried by each individual training sample. Consequently, feature correction methods have also gained widespread attention. In recent years, Pang et al. introduced the Variation Disentangling Generative Adversarial Network (VD-GAN)[5] and the Disentangling Prototype plus Variation model (DisP+V)[6], which generate image centers and feature centers of the training images separately.

3 Method

3.1 Design of Network Structure

To address the issue of attribute interference in single-sample training data, this paper introduces a method based on attribute disentanglement and adversarial generation. It employs a attribute disentanglement network to separate identity information and attribute information within deep facial features and introduces a attribute disentanglement loss to measure the degree of information separation. Simultaneously, a generative adversarial network is constructed to generate features based on the disentangled identity features, enhancing the robustness of the classification network to intra-class variations in facial images. The network architecture is illustrated in Figure 1.

3.2 Pretrained Attribute Disentanglement

The pre-trained attribute disentanglement module consists of four components: feature separation, identity feature classification, attribute disentanglement, and attribute classification. It separates the original features into identity and attribute features. The identity features are used for facial identity recognition,

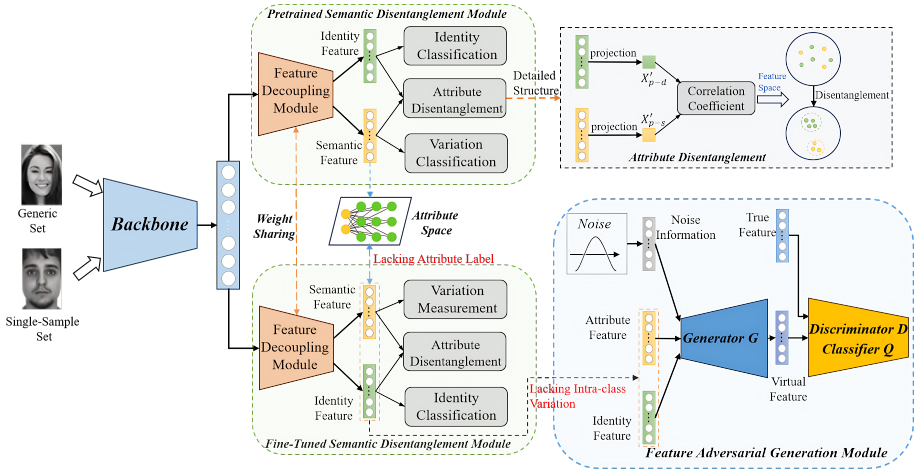

Fig. 1. Framework of our Method. **Step 1**: Train the backbone network and attribute disentanglement module by using the CelebA dataset as a generic set. **Step 2**: Use the attribute features disentangled from the generic set to construct the SOM attribute space, and align single sample attribute features with the attribute space. **Step 3**: Generate feature to increase intra-class variations.

while the attribute features contain information related to facial variations in the image. A attribute disentanglement loss is then used to minimize the correlation between these features by utilizing the correlation coefficient. This loss term enhances facial identity recognition and reduces the impact of facial changes.

Deep Feature Extraction and Disentanglement We employ FaceNet as the deep feature extraction network to map facial images to a deep feature space, represented as $\mathrm{x} = \mathrm{FaceNet}(\mathrm{I}_m)$, where $x \in i^{\,1\times512}$represents the image features, and I_m represents the original image. Based on the principles of the 'P+V' model at the feature level, image features x simultaneously contain identity-related features and attribute features related to variations. So we used a feature disentanglement module to separate identity features $\mathrm{x}_d \in \mathbb{R}^{\,1\times512}$ and attribute features $\mathrm{x}_s \in \mathbb{R}^{\,1\times512}$, as shown in Figure 1.

Identity and Attribute Classifiers After obtaining the identity features of the facial image, denoted as $\mathrm{x}_d \in \mathbb{R}^{\,1\times512}$, the aim is to map the identity information into the feature category space using a deep neural network. To achieve this, we use the Arcface loss, which increases the angular margin between classes, as the loss function for the identity classifier. The Arcface loss can be expressed as follows:

$$L_{id} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\left(\theta_{y_i}+m\right)}}{e^{s\cos\left(\theta_{y_i}+m\right)}+\sum_{j\neq y_i}e^{s\cos\theta_j}} \tag{1}$$

Where N represents the number of classification categories, y_i represents the category label. If we view class probabilities as angles within a full circle, then m can be interpreted as the penalty factor for the angular distance between two categories, and s represents the radius of this circle, which corresponds to the regularized feature.

For acquired attribute features $x_s \in \mathbb{R}^{1\times 512}$, due to intricate facial image variations, this section employs CelebA dataset pretraining. This dataset includes annotated attribute attributes, used to pretrain the network. 17 discriminative attribute labels from this dataset serve as label data for the attribute classifier. To counter noise, noise label classification is introduced during attribute classification, assuming noise in every facial image. This leads to 18 attribute labels. Given each image's noise label, the data's noise part lacks facial information; this label is set to 1 in one-hot encoding. The final loss for the attribute classifier can be expressed as:

$$L_a = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(f\left(x_s\right)\right) \quad = -\frac{1}{N}\sum_{i=1}^{N} \frac{e^{s\cos\left(\theta_{y_i}\right)}}{e^{s\cos\left(\theta_{y_i}\right)} + \sum_{j\neq y_i} e^{s\cos\left(\theta_j\right)}} \tag{2}$$

Where N represents the number of attribute categories, set to 18.

Attribute Disentanglement Loss After seprating identity features x_d and attribute features x_s, their correlation acts as regularization to aid feature disentanglement. Batches of facial images are employed to compute correlations, mitigating feature randomness. Therefore, batch identity features are represented as $X_{p-d} = [x_{d1}, x_{d2}, ..., x_{dn}]$, and batch attribute features are represented as $X_{p-s} = [x_{s1}, x_{s2}, ..., x_{sn}]$, where n represents the number of features. For ease of calculation, batch identity and attribute features are mapped to one-dimensional vectors:

$$X_{p-d}^{'} = W_d^T X_{p-d}, \quad X_{p-s}^{'} = W_s^T X_{p-s} \tag{3}$$

Where, W_d^T and $W_s^T \in \mathbb{R}^{L\times 1}$ represent mapping matrices that map X_{p-d} and X_{p-s} to X'_{p-d} and $X'_{p-s} \in \mathbb{R}^{1\times n}$, respectively. At this point, vectors $X_{p-d}{}'$ and $X'_{p-s} \in \mathbb{R}^{1\times n}$ represent identity and attribute features within this batch. The calculation of the correlation coefficient between these two is as follows:

$$\rho\left(X'_{p-d}, X'_{p-s}\right) = \frac{\operatorname{cov}\left(X'_{p-d}, X'_{p-s}\right)}{\sqrt{D\left(X'_{p-d}\right)}\sqrt{D\left(X'_{p-s}\right)}} \tag{4}$$

Since the covariance between two independent random variables is zero, ρ can be used as a loss term for measuring correlation to reduce the correlation between identity and attribute features. For ease of calculation, ρ^2 is used as the correlation loss and can be expressed as:

$$L_c = \rho^2\left(W_d^T X_{p-d}, W_s^T X_{p-s}\right) \tag{5}$$

Finally, the sum of the identity category loss, attribute category loss, and identity-attribute correlation loss mentioned above is computed as the ultimate network loss:

$$L_{loss} = \alpha L_{id} + \beta L_a + \gamma L_c \tag{6}$$

Where α, β, and γ represent the respective coefficients for the loss terms.

3.3 Fine-Tuned Attribute Disentanglement

Transitioning to single-sample tasks, lacking attribute labels hampers proper attribute classifier usage. To address this, the CelebA dataset serves as a generic set for creating a attribute space with its features. Then, we calculate the distance between the attribute features and the feature spaces. Recognizing human visual systems' advantage in complex information processing, we used a Self-Organizing Map (SOM) network to construct the attribute space, shown in Figure 1. Assuming the set of attribute features separated from the generic set is $X_s^c = [x_s^{c1}, x_s^{c2}, ..., x_s^{cM}]$, where M represents the number of attribute features in the generic set, X_s^c is used as input to construct the SOM network S_c as the attribute space.

For a single-sample image feature x^t, identity features $x_d^t = \mathrm{F}_{split}(\mathrm{x}^t)$, and attribute features $\mathrm{x}_s^t = \mathrm{x}^t$ - $\mathrm{F}_{split}(\mathrm{x}^t)$ are computed using F_{split} network. Identity features x_d^t are trained with the identity classifier in Section 3.2. For attribute features x_s^t, Mean Square Error (MSE) loss measures the distance to attribute space, facilitating attribute feature training. The attribute loss can be expressed as:

$$L_{mse} = \frac{1}{m} \sum_{i=1}^{m} \left(x_s^{ti} - n_c^i \right) \tag{7}$$

Where, m is feature dimension, x_s^{ti} is the i-th dimension of attribute feature x_s^t, and n_c^i is the i-th dimension of winning neuron for x_s^t in attribute space.

3.4 Feature Adversarial Generation

In single sample face recognition, limited training samples and diverse test samples cause disentanglement networks to deviate from true category centers. Misclassification is common due to small intra-class distances and sample variations. To address this, we propose using obtained identity features as category centers and generating virtual features. This improves classification margin and mitigates center deviations.

Unlike traditional GANs, infoGAN uses real features and noise for interpretable attribute generation. The mutual information constraint between generated features and input information ensures the interpretability of the generation process. It includes a feature generator (G), discriminator (D), and category classifier (Q) for authenticating features and categories.

Using the attribute space S_c constructed in Section 3.3, the identity feature x_d is used as input, and the obtained winning neurons are used as attribute

feature x_s. The generated attribute feature x_s, disentangled identity feature x_d, and noise information are sampled and input into the Generative Adversarial Network, represented as follows:

$$\hat{x} = G\left(f\left(x_d\right)\right) = G\left(x_d + x_s + W_c c\right) \tag{8}$$

Where $\hat{x}$ is the generated image feature, W_c are noise weights, and $c \in \mathbb{R}^{1\times 512}$ is Gaussian noise, simulating diverse facial image variations.

The use of randomly generated Gaussian noise can generate numerous intra-class variation images, denoted as $\hat{X} = [\hat{x}^0, \hat{x}^2, \ldots, \hat{x}^C]$, where C represents the number of virtual features generated for the same class, set to 100.

Here, feature classifier Q measures the correlation between undisentangled real feature x and virtually generated feature $\hat{x}$. Both maintain identical identity labels before and after generations. The identity classifier ensures consistent labeling, preserving identity information during feature generation. This method prioritizes identity preservation, boosting classification efficiency over pre- and post-generation mutual information calculations.

The final feature generation network loss can be expressed as:

$$\begin{aligned} \text{Loss} = & E_{x\sim p_z(z)}\left[\log\left(1 - D\left(G\left(x_d + x_s + W_c c\right)\right)\right)\right] \\ & + E_{x\sim p_{\text{data}}(x)}[\log D(x)] - \lambda[\text{soft}\max(x) + \text{softmax}(\hat{x})] \end{aligned} \tag{9}$$

4 Experimental Results

To validate our method's efficacy in single-sample face recognition, we conducted experiments, comparing them with state-of-the-art techniques. The AR [18], LFW [8], and FERET [19] datasets were employed, with FaceNet extracting 512-dimensional deep facial features. The disentanglement network pre-training utilized the CelebA dataset[7] to enhance identity and attribute features separation. In training, a attribute classifier used 17 selective variation labels, and images were resized to 160x160. As in prior works[13], we adopted Accuracy (acc) as our metric.

4.1 Results on AR Dataset

In this section, we conducted experiments using 100 classes from the AR dataset. 80 classes were used for training and testing in single-sample, while the remaining 20 classes were used as a generic set. The first image of each class was used as the training sample. We compared our proposed method with commonly used algorithms for single-sample face recognition, including traditional and deep learning methods. The results are shown in Table 1.

According to the experimental results, our proposed method achieved a minimum 5% improvement compared to traditional methods like AGL, RHDA, and SVDL. Additionally, our proposed method achieved a 0.53% improvement over other deep learning methods like SSLD and KWV. These results highlight the robustness of our proposed method in handling facial variations such as expressions, lighting, and occlusions.

Table 1. AR Dataset Comparison Experiment

Method	Acc	Method	Acc
AGL	59.95	SSAE	85.21
BlockFLD	62.92	SSLD	94.67
ESRC	71.88	SSPP-DAN	93.33
SVDL	72.56	KWV	94.31
RHDA	90.65	VD-GAN	79.70
SGL	87.30	FaceNet	86.30
FDDL	95.00	**ours**	**95.20**

4.2 Results on LFW Dataset

Compared to AR, LFW contains a greater amount of unconstrained variation information, including over 13,000 facial images from more than 5,000 categories. We used 158 classes from LFW for training and testing in single-sample recognition, and 1522 classes for the generic set. Results are in Table 2.

Table 2. LFW Dataset Comparison Experiment

Method	Acc	Method	Acc
AGL	31.90	SSPP-DAN	97.91
BlockFLD	18.10	Center-Fea	90.60
ESRC	33.60	CJR-RACF	95.50
SVDL	33.50	DisP+V	96.70
VQ	42.92	UP	94.80
RHDA	32.90	FaceNet	93.80
SSLD	92.70	**ours**	**98.34**

Our proposed method achieved a high classification accuracy of 98.34% even with complex and unconstrained facial variations. Compared to other deep learning methods, our proposed method showed a minimum of 0.43% improvement in accuracy. In comparison to the DisP+V method of generating virtual features, our proposed method still achieves a 1.64% improvement.

4.3 Results on FERET Dataset

In this section, we used the FERET-b dataset, where 200 classes were used for experimental validation, with each class containing 7 intra-class samples. The experimental results are shown in Table 3.

Our proposed method outperformed traditional methods like SRC, ESRC, CPL, and deep learning methods like SSPP-DAN and KCFT by at least 5.4%. These results highlight the effectiveness of our method in achieving high classification accuracy, even in the presence of common facial variations such as lighting and pose changes.

Table 3. FERET Dataset Comparison Experiment

Method	Acc	Method	Acc
SRC	53.44	SSPP-DAN	93.30
ESRC	58.90	KCFT	93.17
CPL	93.67	FaceNet	91.40
TDL	89.33	**ours**	**99.30**

4.4 Ablation Study

To validate the effectiveness of our method, this section conducted ablation experiments on two modules to compare the accuracy on different datasets. The experimental results are shown in Table 4.

Table 4. Ablation on Attribute Disentanglement and Feature Augmentation Module

Attribute Disentanglement	Feature Augmentation	Acc		
		AR	LFW	FERET
		86.3	93.8	91.4
✓		92.2	96.8	98.3
✓	✓	**95.2**	**98.3**	**99.3**

Experimental results reveal the substantial enhancements from the proposed attribute disentanglement and featrue augmentation modules compared to relying solely on FaceNet-extracted features. The disentanglement and augmentation modules achieved 5.9% and 3% improvements on the AR dataset, respectively. The disentanglement module alone achieves 92.2% accuracy in AR's natural images. Despite LFW dataset variations, the augmentation module outperforms the disentanglement module, adding 1.5% accuracy.

Ablation experiments confirm the disentanglement module's effectiveness in mitigating attribute variations' impact on accuracy, while the featrue augmentation module enhances accuracy against common and complex scenarios.

5 Conclusion

This study addresses the challenge of balancing visual and attribute information in single-sample face recognition. We propose an approach using attribute disentanglement and adversarial augmentation. Our method employs a feature disentanglement network to separate identity from attribute features. Utilizing identity features as category centers, a generative adversarial network creates virtual features, enhancing classification accuracy. Experimental results on AR, LFW, and FERET datasets show our method maintains strong classification performance under complex facial variations, effectively distinguishing intra-class and inter-class differences.

References

1. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE/CVF international conference on computer vision. (2019) 9459–9468
2. Tran, L., Yin, X., Liu, X.: Representation learning by rotating your faces. IEEE transactions on pattern analysis and machine intelligence **41**(12) (2018) 3007–3021
3. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Feature transfer learning for face recognition with under-represented data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 5704–5713
4. Min, R., Xu, S., Cui, Z.: Single-sample face recognition based on feature expansion. IEEE Access **7** (2019) 45219–45229
5. Pang, M., Wang, B., Cheung, Y.m., Chen, Y., Wen, B.: Vd-gan: A unified framework for joint prototype and representation learning from contaminated single sample per person. IEEE Transactions on Information Forensics and Security **16** (2021) 2246–2259
6. Pang, M., Wang, B., Ye, M., Chen, Y., Wen, B.: Disentangling prototype and variation for single sample face recognition. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE (2021) 1–6
7. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. (2015) 3730–3738
8. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. Month (2008)
9. Yang, M., Wang, X., Zeng, G., Shen, L.: Joint and collaborative representation with local adaptive convolution feature for face recognition with single sample per person. Pattern Recognition **66**(C) (2016) 117–128
10. Duan, Q., Zhang, L.: Look more into occlusion: Realistic face frontalization and recognition with boostgan. IEEE Transactions on Neural Networks and Learning Systems **PP**(99) (2020) 1–15
11. Zhou, J., Chen, J., Liang, C., Chen, J.: One-shot face recognition with feature rectification via adversarial learning. (2020)
12. Pang, M., Cheung, Y.M., Wang, B., Lou, J.: Synergistic generic learning for face recognition from a contaminated single sample per person. IEEE transactions on information forensics and security **15**(1) (2020) 195–209
13. Wang, X., Zhang, B., Yang, M., Ke, K., Zheng, W.: Robust joint representation with triple local feature for face recognition with single sample per person. Knowledge-Based Systems **181** (2019) 104790–
14. Liu, F., Chen, D., Wang, F., Li, Z., Xu, F.: Deep learning based single sample face recognition: a survey. Artificial Intelligence Review: An International Science and Engineering Journal (2023)
15. Kohonen, T.: The self-organizing map. IEEE Proc Icnn **1**(1-3) (1990) 1–6
16. Zhuo, T.: Face recognition from a single image per person using deep architecture neural networks. Cluster Computing **19**(1) (2016) 73–77
17. Choe, J., Park, S., Kim, K., Park, J.H., Kim, D., Shim, H.: Face generation for low-shot learning using generative adversarial networks. In: IEEE International Conference on Computer Vision Workshop. (2017) 1940–1948
18. Martinez, A., Benavente, R.: The ar face database: Cvc technical report, 24. (1998)
19. A, P.J.P., B, H.W., B, J.H., A, P.J.R.: The feret database and evaluation procedure for face-recognition algorithms. Image and Vision Computing **16**(5) (1998) 295–306