2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Group-wise Feature Orthogonalization and Suppression for GAN based Facial Attribute Translation

Zhiwei Wen, Haoqian Wu, Weicheng Xie*, Linlin Shen
Computer Vision Institute, Shenzhen University, Shenzhen, China
Shenzhen Institute of Artificial Intelligence and Robotics for Society, PR China
Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, PR China
Email: {wenzhiwei2018,wuhaoqian2019}@email.szu.edu.cn, {wcxie,llshen}@szu.edu.cn

***Abstract*—Generative Adversarial Network (GAN) has been widely used for object attribute editing. However, the semantic correlation, resulted from the feature map interaction in the generative network of GAN, may impair the generalization ability of the generative network. In this work, semantic disentanglement is introduced in GAN to reduce the attribute correlation. The feature maps of the generative network are first grouped with an efficient clustering algorithm based on hash encoding, which are used to excavate hidden semantic attributes and calculate the group-wise orthogonality loss for the reduction of attribute entanglement. Meanwhile, the feature maps falling in the intersection regions of different groups are further suppressed to reduce the attribute-wise interaction. Extensive experiments reveal that the proposed GAN generated more genuine objects than the state of the arts. Quantitative results of classification accuracy, inception score and FID score further justify the effectiveness of the proposed GAN.**

I. Introduction

Since the introduction of generative adversarial network (GAN) by Goodfellow [1], it has achieved impressive performance for attribute editing of various generated problems. When a GAN learns from a large training dataset intensively, the feature maps are easily to be highly entangled with each other. However, highly entangled information learned from the training dataset is hard to be applied to the testing dataset due to large entanglement variance between different identities and datasets. Meanwhile, the entangled information makes it hard to edit one attribute independently since the entanglement may influence the generation of other attributes.

To reduce the correlation among different feature maps, various network deployments and loss functions were proposed. Yang and Tseng [2] proposed to construct the handcrafted orthogonal basis function in a network layer, while the number of network layers and the convergence speed are largely limited when the number of orthogonal functions is large. Chen et al. [3] proposed to maximize the mutual information entropy of the network feature and latent variable representation to reduce the feature entanglement for object synthesis. Rifai et al. [4] employed contractive discriminative analysis to separate the emotion-related factors from others. Shaham et al. [5] imposed orthogonality on the pair-wise output. Wan et al. [6] employed Gram-Schmidt orthogonalization to reduce the interference among the nodes to improve network generalization ability. Hong et al. [7] reviewed the algorithms for latent space decomposition for the generative network of GANs.

However, the node-wise or feature map-wise disentanglement is inefficient for the network with large number of nodes or feature maps. Meanwhile, the semantic attribute of an object is often encoded with a bunch of network hidden outputs [8], feature map-wise operation is hard to capture the information about semantic attribute. Semantic attributes were often excavated and pruned, while their disentanglement enables the independent editing of each attribute and better generalization ability of the network to the testing dataset.

To explore the semantic attributes based on feature maps, Klemmer et al. [9] employed the local correlation of the feature neighborhood during the conditional generation. Wu and He [10] uniformly divided the feature maps into groups and performed group-wise normalization to simulate attribute-wise regularization. Chen et al. [11] proposed to suppress each pair of two orthogonal groups with a regression loss based on the segmentation of foreground and background regions. Wang et al. [12] introduced the clustering based on the appearance of the feature maps to prune the filters, compress and accelerate CNNs. Mukherjee et al. [13] used the latent feature representation of GAN for sample clustering. Kazemi et al. [14] decomposed the latent feature representation into the content and style codes to encode the geometrical and textural information in a generative network. Luan et al. [15] disentangled the head pose from the identity representation by embedding the pose code into the decoder, which enables the network to synthesize a face with a given identity and the target pose. Shu et al. [16] disentangled the latent feature vector into the representations of the texture and geometrical deformations, and then used them for image reconstruction. By orthogonalizing the hidden vectors based on the normal direction of latent feature projection, Shen et al. [17] allowed network to edit one semantic attribute independently. Liang et al. [8] employed the clustering of the fully connected layer in the discriminator network to not only enable the disentanglement of discriminator and generator, but also explore the

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412434

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on August 13,2021 at 08:59:02 UTC from IEEE Xplore. Restrictions apply.

latent attributes.

From another aspect, feature maps can easily contain noisy or redundant information after training. Hence the pruning of feature maps is often considered to reduce network complexity and improve generalization ability. To decrease feature redundancy and network computation, intrinsic representation of feature maps is employed with feature discriminability preservation [18]. He et al. [19] revealed that the pruning of layer channels can not only accelerate network computation, but also yield better generalization performance on small dataset. Ayinde et al. [20], [21] proposed to prune the feature maps randomly located in each group to eliminate redundant features after clustering. Tian et al. [22] showed that the outlier sample elimination during network training is beneficial to the improvement of recognition performance. Tompson et al. [23] proposed to drop out feature maps entirely in a discriminative network to improve its generalization ability.

A. Motivation

While the correlation of feature maps are explored in the previous works, the latent semantic attributes implied in the feature map groups are rarely studied. In order to excavate semantic attributes, the feature maps during the face generation are grouped into three clusters and their averages are demonstrated in the 1st row of Fig. 1. As shown in Fig. 1, each group of feature maps mainly responds to one of the semantic attributes, i.e., face, hair or background. This observation motivates us to disentangle these feature maps to enable editing of each attribute independently during facial attribute translation.

From another aspect, it is shown in previous generative networks that the suppression of outlier feature maps (FMs) is beneficial for network generalization capability. To study the effect of outlier feature map (FM) suppression for facial attribute translation, we demonstrated an example hair color translation with and without outlier FM suppression in the 2nd row of Fig. 1. Fig. 1(i) shows that the abnormally generated color in the red rectangle is largely improved in Fig. 1(g) by eliminating outlier FM in Fig. 1(f). This observation motivates us to suppress outlier or noisy feature maps like Fig. 1(f), while increase the contribution of normal feature maps like Fig. 1(h), during the training of attribute translation.

B. Contributions

In this work, we propose to suppress feature maps in terms of groups that reflect the semantic attributes. The group-wise feature map reduction can decrease the redundant information in the trained model. Meanwhile, the suppression of features that fall in the intersection regions of groups can further disentangle the attribute interactions. Thus, feature group-wise orthogonalization and intersection suppression are proposed to reduce semantic attribute entanglement. The feature maps are first grouped with an efficient clustering to excavate semantic attributes, a group-wise orthogonality loss is then followed to enable independent attribute editing. The feature maps are further group-wise suppressed to reduce attribute interaction.

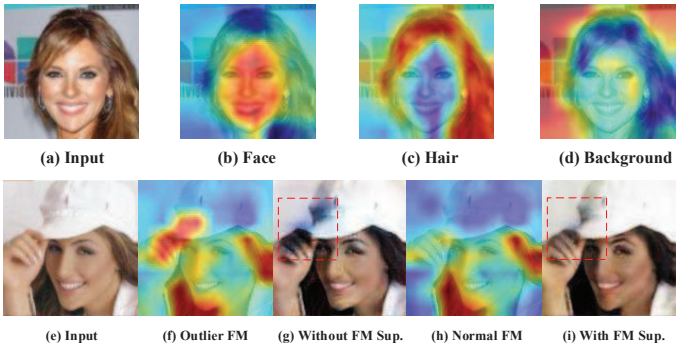

Fig. 1. (b)-(d): The average of grouped feature maps corresponding to semantic attributes of face, hair and background, respectively; (g), (i): The generated images of hair color translation from blond to black without and with the suppression (Sup.) of the feature map (FM) in (f), where red rectangles labeled their main difference.

Fig. 2. The input and output of hair color translation of $StarGAN$, $AttGAN$ and ours. The dotted red rectangle labeled the main difference between the baseline synthesis and ours.

The performances of the proposed GAN are compared with the baselines, i.e. $StarGAN$ [24] and $AttGAN$ [25], and demonstrated in Fig. 2.

The main contributions of the proposed algorithm are presented as follows

- Feature map clustering is proposed to excavate semantic attributes and reduce their correlation based on an effective encoding of the feature maps, i.e. hash encoding;
- Two approaches, i.e. group-wise orthogonality and intersection feature suppression are proposed to reduce attribute interaction;
- Better generalization performances on visual and quantitative results are achieved by the proposed GAN.

This paper is structured into the following sections. The proposed approach is introduced in Section II. The experimental results and the corresponding illustrations are demonstrated in Section III. Finally, the conclusions and a discussion are presented in Section IV.

II. The Proposed GAN

The framework of the proposed GAN is presented in Fig. 3. The feature maps are first grouped based on an efficient hash

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on August 13,2021 at 08:59:02 UTC from IEEE Xplore. Restrictions apply.

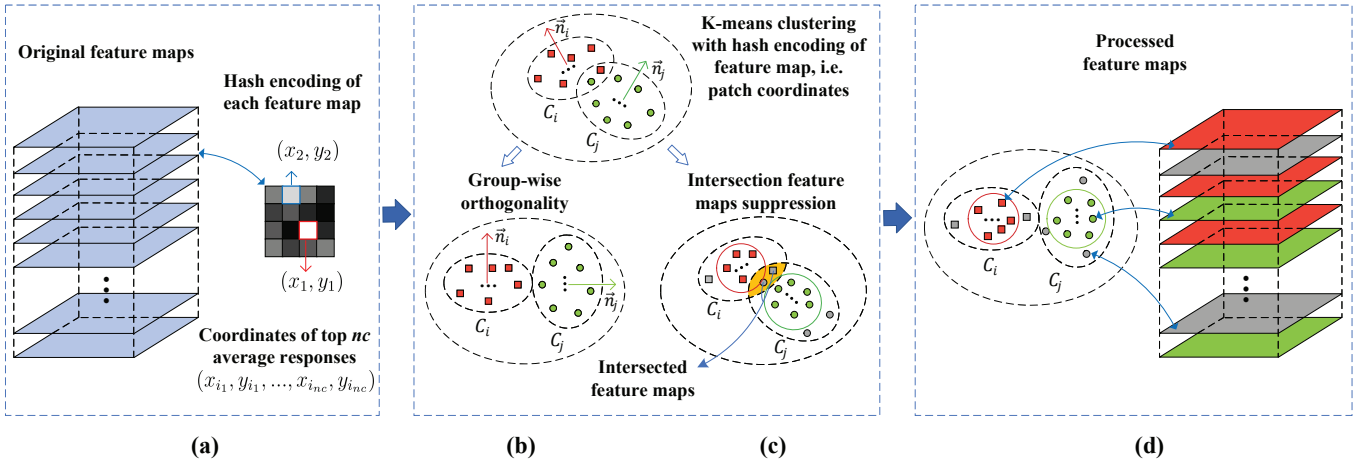

Fig. 3. The framework of the proposed GAN. Each feature map is hash-encoded with the coordinates of top nc average responses (a). The efficient clustering, group-wise orthogonality (b) and intersection suppression (c) are then followed to reduce the entanglement of semantic attributes. The green circles and red rectangles denote the feature maps from two groups, while $\vec{n_i}$ and $\vec{n_j}$ are the corresponding normal directions. Processed feature maps are shown in (d), where the red and green maps are transformed with an orthogonality loss and the gray maps are suppressed during training.

encoding. The group-wise orthogonality and intersection feature suppression are then followed to reduce the entanglement of semantic attributes for the generative networks.

A. Feature Map Encoding and Clustering

For hidden attribute excavation, k-means algorithm is employed to cluster feature maps. However, the runtime cost can be high when the dimension of the vectorized feature map is large.

As an alternative, the feature maps are clustered according to their hash encoding, i.e. the coordinates of several top largest responses of the segmented patches. More precisely, each feature map with the size of 32×32 is first divided into 8×8 patches, i.e. each patch contains 4×4 pixels. The coordinates of the patches with the top nc average responses are used for the clustering, which are vectorized as follows

$$t = (x_{i_1}, y_{i_1}, \cdots, x_{i_{nc}}, y_{i_{nc}}) \tag{1}$$

where (x_{i_j}, y_{i_j}) is the coordinate of the patch in 8×8 patches with the top j-th largest average response, the setting of $nc = 2$ is employed.

In order to study the performance of the employed hash encoding, feature encoding with principal component analysis (PCA) [26] and direct vectorization, are used for a comparison. The runtime costs of both feature map encoding and the corresponding k-means clustering are taken into account. The time complexities of hash encoding in equation (1) and PCA are $O(N \cdot n)$ and $O(N \cdot n) + O(n^3)$, where $N = 256$ and $n = 1024 = 32 \times 32$ are the number and dimension of the vectorized feature maps. Since $O(n^3) >> O(N \cdot n)$, the time complexity of PCA encoding, i.e. $O(N \cdot n) + O(n^3)$, largely exceeds the runtime cost of the proposed hash encoding, i.e. $O(N \cdot n)$. The runtime cost of direct vectorization is almost negligible compared with PCA and hash encodings. Meanwhile, the time complexity of k-means is $O(nIter \cdot N \cdot m \cdot n)$, where $nIter$ and m are the numbers of iterations and clusters. When PCA and hash encodings are employed, n decreases from $1024 = 32 \times 32$, i.e. the size of original feature map, to $2 \cdot nc$ ($nc = 2$). Thus, the runtime costs of k-means clustering with PCA and hash encoding of feature maps are largely reduced compared with that of the direct vectorization. Consequently, the feature map clustering based on hash encoding demands the least runtime cost for the network training.

B. Group-wise Feature Orthogonality and Suppression

While the feature maps are clustered into m groups, the group-wise orthogonality loss is formulated as follows

$$\mathcal{L}_{GO} = \frac{2}{m(m-1)} \Sigma_i \Sigma_j g_i^T g_j, \tag{2}$$

where m is the number of clusters, g_i, g_j are the average feature maps in the i-th and j-th groups and formulated as follows

$$\begin{cases} g_i = \Sigma_{k \in C_i} w_k^{(i)} f_k, \\ ||f_k||_2 = \gamma, \end{cases} \tag{3}$$

where C_i records the i-th group of feature maps, $w_k^{(i)}$ is the weight value corresponding to the feature map f_k in the i-th group C_i satisfying $\sum_k w_k^{(i)} = 1$, γ is a normalization parameter optimized in the study [27]. In this work, uniform weight values, i.e. $w_k^{(i)} = 1/\#C_i$, is employed, while more profound weighting method with self attention or correlation attention [28] can be considered for improvement.

The group-wise orthogonality can not only decrease the runtime cost of the feature map-wise orthogonality, but also disentangle the correlation between hidden semantic attributes.

Since the feature maps located far away from the average of grouped features are more likely to fall in the intersection regions of feature groups, they are selected and suppressed during the network training.

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on August 13,2021 at 08:59:02 UTC from IEEE Xplore. Restrictions apply.

Motivated from the study [22], the distance between two feature maps, i.e. f_i and f_j, obeys a normal distribution as follows

$$||f_i - f_j||_2 \sim \mathcal{N}(\sqrt{2}\gamma, \frac{\gamma}{\sqrt{2n}}), \tag{4}$$

where γ is the L_2-norm of the feature map embedded with the encoder of a generative network; n is the dimension of the vectorized feature map. Then the distance between each feature map and the corresponding average, i.e. d_i, should obeys the following distribution

$$d_i = ||f_i - \frac{1}{N^{(r)}} \sum_j f_j||_2 \sim \mathcal{N}(0, \frac{\gamma}{\sqrt{2n}N^{(r)}}), \tag{5}$$

where $N^{(r)}$ denotes the number of samples in the r-th group, i.e. $N^{(r)} = \#\{C_r | f_i \in C_r\}$. Feature map f_i is selected to be suppressed if it deviates from the average feature map with a distance of κ times the variance as follows

$$d_i > \kappa \frac{\gamma}{\sqrt{2n}N^{(r)}}, \tag{6}$$

where $\kappa = 1.5$ is a predetermined value.

As shown in Fig. 3, the feature maps detected according to Eq. (6) are further suppressed during the training. In order to offset the effect of the feature maps suppressed in the intersection regions, a reconstruction bias with and without the intersection feature suppression is applied on the generator network to preserve the genuineness of the attribute translation.

C. Network Training

The proposed group-wise orthogonality loss and intersection feature suppression are embedded in a GAN for attribute editing.

GAN [1] consists of a generator G and a discriminator D in a two-player minimax game. Based on the proposed group-wise orthogonality loss and intersection feature suppression on the generative network, final network losses of the proposed GAN are formulated as follows

$$\begin{cases} \mathcal{L}_{drop} = E_{x,c'}[||G(x,c') - G_{drop}(x,c')||_1], \\ \mathcal{L}_{rec} = E_{x,c,c'}[||x - G_{drop}(G(x,c'),c)||_1], \\ \mathcal{L}_G = \mathcal{L}_{OriG} + \lambda_{drop}\mathcal{L}_{drop} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{GO}\mathcal{L}_{GO}, \\ \mathcal{L}_D = \mathcal{L}_{OriD}. \end{cases} \tag{7}$$

where G generates an image $G(x,c')$ conditioned on both the input image x and target domain label c', G_{drop} denotes the generator with intersection feature dropout and c is original domain label. $\mathcal{L}_{drop}$ is a loss reflecting the bias of generator with and without the proposed intersection feature dropout to enable GAN to suppress the effect of the feature maps in the intersection regions. Meanwhile, a reconstruction loss of $\mathcal{L}_{rec}$ is applied on the generator to preserve reconstruction genuineness after intersection feature dropout. $\mathcal{L}_{OriG}$ and $\mathcal{L}_{OriD}$ denote the loss functions of the original generator and discriminator networks, and λ_{drop}, λ_{rec} and λ_{GO} are regularization hyper-parameters. Since the contribution of intersection features in the network is reduced by the defined losses of $\mathcal{L}_{drop}$ and $\mathcal{L}_{rec}$, intersection feature suppression is achieved. While the feature map are dropout only for generator network training to suppress outlier feature maps, these feature maps are still active in the testing network for attribute translation.

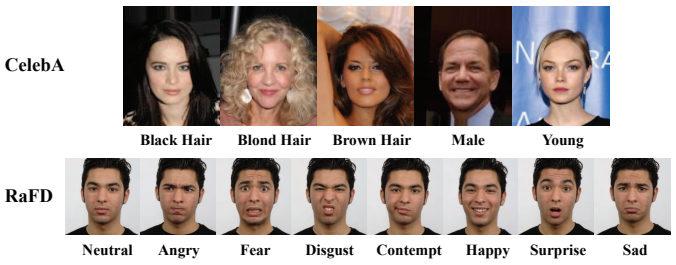

Fig. 4. The example images of CelebA [29] and RaFD [30].

In this work, the $StarGAN$ and $AttGAN$ are used as the baselines for comparison, hence $\mathcal{L}_{OriG}$ and $\mathcal{L}_{OriD}$ are the losses of the generator and discriminator of $StarGAN$ or $AttGAN$. Finally, the objective functions of the generator G and the discriminator D are $\mathcal{L}_G$ and $\mathcal{L}_D$, respectively. Based on the proposed losses, the similar minimization-maximization algorithm as $StarGAN$ or $AttGAN$ is performed.

For the convenience of the following presentation, the proposed GAN is abbreviated as $GAN_x + GO + IFS$, where GAN_x denotes $StarGAN$ or $AttGAN$. GO and IFS are the abbreviations of the proposed group-wise orthogonality loss and intersection feature suppression, respectively.

III. Experimental Results

Two databases are used for the testing, i.e. CelebFaces Attributes Dataset (CelebA) [29] and Radboud Faces Database (RaFD) [30]. CelebA is a large-scale face attributes dataset with more than 200K celebrity images, which are annotated with 40 attributes. While the top 20K images of CelebA are used for our experiment, 19K of which are selected for training and the rest are used for testing. RaFD contains pictures of 67 models displaying 8 emotional expressions, i.e., angry, fear, disgust, contempt, happy, surprise, sad and neutral. For the RaFD database, 4320 and 504 images are used for training and testing, respectively. Example images of the two employed databases are shown in Fig. 4.

We performed the experiments using Pytorch platform and the same network structure as $StarGAN$ [24] [1] or $AttGAN$ [25] [2]. All the models are trained for 40,000 steps. The setting of the regularization parameters, i.e. $\lambda_{drop} = 15$, $\lambda_{rec} = 3$ and $\lambda_{GO} = 5$ in equation (7), are employed. The number of k-means clusters, i.e. m in Eq. (2), is set as 5 and 8 for the databases of CelebA and RaFD, respectively. More precisely, the number of clusters is set as the number of attributes to allow a group of feature maps to respond to an attribute.

[1] https://github.com/yunjey/StarGAN
[2] https://github.com/elvisyjlin/AttGAN-PyTorch

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on August 13,2021 at 08:59:02 UTC from IEEE Xplore. Restrictions apply.

TABLE I
THE AVERAGE ACCURACIES, INCEPTION SCORE (IS), FID AND RUNNING TIME (RT, IN SECONDS) OF HASH ENCODING, PCA AND DIRECT VECTORIZATION BASED ON $StarGAN + GO + IFS$ FOR RAFD.

Method	Accuracy ↑	IS ↑	FID ↓	RT ↓
Hash encoding	98.41%	**2.770**	**43.51**	**0.102**
PCA	98.21%	2.769	44.71	0.261
Direct vectorization	**98.81**%	2.759	46.92	1.289

↑ means larger numbers are preferred, ↓ means opposite.

A. Hash Encoding

Section II-A presents a theoretical analysis of the runtime cost of the proposed hash encoding compared with PCA and direct vectorization. In this section, we provide an experimental study of the runtime costs of three methods for feature map clustering, including the feature map encoding and the k-means clustering. Meanwhile, the quantitative results with the metrics of average accuracy, inception score (IS) [31] and FID score [32] for the images generated based on the three encodings are presented in Table I. For classification accuracy, Resnet-18 [33] is used to recognize each attribute, while the training set of GAN is used for classifier training, the images generated by the considered GAN variant are used as the testing set. IS and FID scores are used to measure the quality and diversity of the generated images.

We evaluated the average runtime cost with respect to the batch size of 16, and trained $StarGAN + GO + IFS$ for the RaFD dataset. Table I shows that the average runtime cost of the hash encoding, PCA and direct vectorization are 0.102, 0.261 and 1.289 seconds, respectively. The least runtime cost with hash encoding verifies the theoretical analysis in Section II-A. As shown in Table I, hash encoding performs similarly compared with PCA and direct quantization in terms of classification accuracy, while achieves slightly better performances than the other two encodings in terms of IS and FID scores. Therefore, hash encoding does not cause large information degradation in terms of classification accuracy, IS and FID scores, while largely decreases the runtime cost of feature map encoding and clustering.

B. Visual Results

To study the performance of the group-wise orthogonalization, the proposed loss is compared to the feature map-wise orthogonalization. To avoid large model complexity introduced by pairwise orthogonality of feature maps, an additional layer is appended in the generative network [28] to simulate the correlation matrix of the feature maps. Then the feature map-wise orthogonality loss is approximated as the average of the correlation matrix. Fig. 5 shows an example face translated from female to male without any orthogonality loss (b), with the feature map-wise loss (c) and the proposed group-wise orthogonality loss (d). To study the performance of the proposed feature map suppression, the generations with and without the proposed feature suppression for the transfer of brown hair are presented in Fig. 6. The offline feature suppression is performed after network training, while the online suppression is executed during training.

Fig. 5. The comparison of two orthogonality strategies.

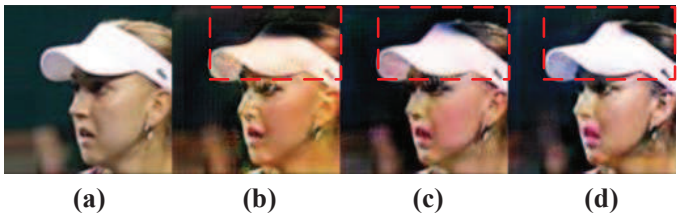

Fig. 6. Hair color translation for a lady face (a) by $StarGAN$ (b), the offline (c) and the online (d) feature suppression.

Fig. 7. Generated expressions translated by $StarGAN$, $StarGAN$+IFS, $StarGAN$+GO and $StarGAN$+IFS+GO for RaFD.

Fig. 5 shows that the proposed orthogonality strategy can better preserve the texture information of original input during the image generation. Meanwhile, more clean face is generated since the noise is suppressed with the proposed orthogonality loss. Fig. 6(c) reveals that the feature maps suppressed offline can capture the abnormal regions during attribute generation when these feature maps are dropout after training. As shown in Fig. 6(d), when the feature maps are suppression online, the abnormal regions are reduced, which justifies the employed online dropout in the proposed feature map suppression.

To further study the performances of GO and IFS, we performed the ablation study by embedding GO, IFS and IFS+GO into $StarGAN$, then their performances on the RaFD dataset are compared to the baseline, i.e. $StarGAN$, and the results are shown in Fig. 7. Fig. 7 shows that the proposed group-wise orthogonality and intersection feature suppression both improve the performance of expression translation.

Furthermore, the overall performance of the the proposed GAN, i.e. $GANx + GO + IFS$, is compared to that of the corresponding baseline, i.e. $GANx = StarGAN$ or $GANx = AttGAN$, for the CelebA database, and the generations are demonstrated in Fig. 8. Fig. 8 shows that the proposed GAN outperforms $StarGAN$ and $AttGAN$ on

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on August 13,2021 at 08:59:02 UTC from IEEE Xplore. Restrictions apply.

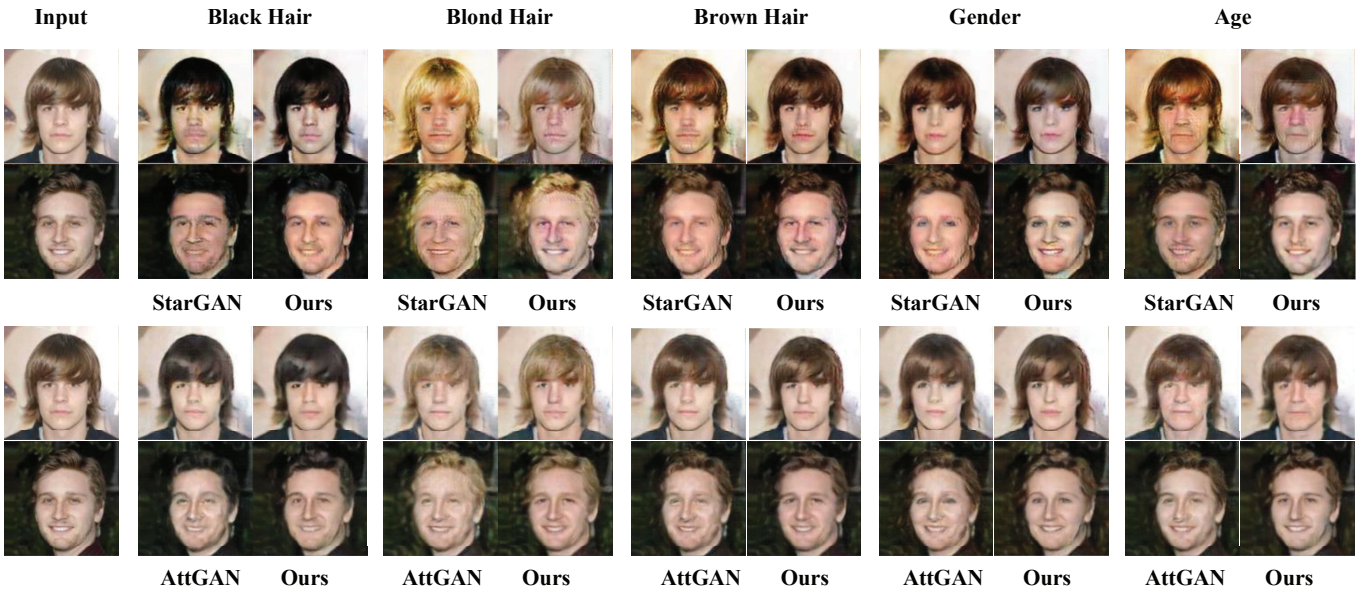

Fig. 8. The performances of $StarGAN$, $AttGAN$ and the proposed GAN for CelebA.

TABLE II
THE ACCURACY (ACC. %), IS AND FID SCORES OF $StarGAN$, $AttGAN$ AND THE PROPOSED GAN ($GAN_x + IFS + GO$) FOR EACH ATTRIBUTE OF CELEBA.

Meas.	Method	Black	Blond	Brown	Gender	Age
Acc. ↑	$StarGAN$	66.77	**78.98**	55.96	60.06	**63.46**
	Ours	**78.08**	71.37	**63.76**	**63.06**	63.36
	$AttGAN$	50.55	31.23	34.33	63.96	58.16
	Ours	**55.16**	**42.14**	**36.04**	**65.97**	**69.37**
IS ↑	$StarGAN$	1.178	1.182	1.014	1.200	1.122
	Ours	**1.204**	**1.237**	**1.033**	**1.221**	**1.127**
	$AttGAN$	1.317	1.329	1.131	1.310	1.111
	Ours	**1.331**	**1.344**	**1.157**	**1.325**	**1.121**
FID ↓	$StarGAN$	65.90	93.06	71.28	105.06	85.58
	Ours	**58.86**	**79.92**	**64.86**	**101.73**	**76.68**
	$AttGAN$	62.29	84.94	66.41	101.03	82.56
	Ours	**56.63**	**82.51**	**60.97**	**98.78**	**78.06**

↑ means larger numbers are preferred, ↓ means opposite.

TABLE III
THE AVERAGE ACCURACIES, IS AND FID SCORES OF DIFFERENT GAN VARIANTS FOR RAFD.

Method	Accuracy ↑	IS ↑	FID ↓
$StarGAN$	97.62%	2.516	46.59
$StarGAN$+IFS	97.02%	2.673	46.53
$StarGAN$+GO	97.62%	2.617	44.56
$StarGAN$+IFS+GO	**98.41%**	**2.770**	**43.51**
$AttGAN$	65.48%	2.785	60.39
$AttGAN$+IFS	69.84%	2.803	51.02
$AttGAN$+GO	70.63%	2.827	**50.01**
$AttGAN$+IFS+GO	**73.02%**	**2.918**	53.24

↑ means larger numbers are preferred, ↓ means opposite.

both genuineness and clearness. Especially in the 2nd row, the semantic disentanglement enables the proposed algorithm to generate more normal textures in the forehead region compared with $StarGAN$ and $AttGAN$.

C. Quantitative Results

To quantitatively evaluate the performance of the proposed GAN, the classification accuracies, IS and FID scores of $StarGAN$, $AttGAN$ and the proposed GAN for different attribute translations on the CelebA dataset are shown in Table II. Table III further presents the ablation study of these methods for the RaFD database.

Table II shows that the proposed GAN outperforms $StarGAN$ and $AttGAN$ for most attributes in terms of inception score and FID score, which reveals that the proposed GAN can synthesize high quality images although there is large variation between the training and testing datasets. For the classification accuracy, the proposed GAN achieves better performances for most attributes, while worse performances for blond hair and age translations. To study the generations with the proposed GAN that are wrongly classified, Fig. 9 presents example generations of the proposed GAN that are wrongly classified, and the generations of $StarGAN$ that are correctly recognized. While the classification accuracy does not necessarily reflect the quality of generated images, one can observe that the proposed GAN achieves more genuine and clean parts, e.g. forehead, than $StarGAN$.

Based on the ablation study of expression transformation for the RaFD dataset, Table III shows that the proposed GAN, i.e. $GAN_x + IFS + GO$, outperforms the both baselines of $StarGAN$ and $AttGAN$ in terms of classification accuracy, IS and FID, and large improvements of 7.54% and 7.15 are achieved for the accuracy and FID scores compared with the baseline of $AttGAN$. Meanwhile, the proposed $GAN_x + IFS + GO$ performs better than the GAN variant with IFS or GO in almost all the cases, which verifies

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on August 13,2021 at 08:59:02 UTC from IEEE Xplore. Restrictions apply.

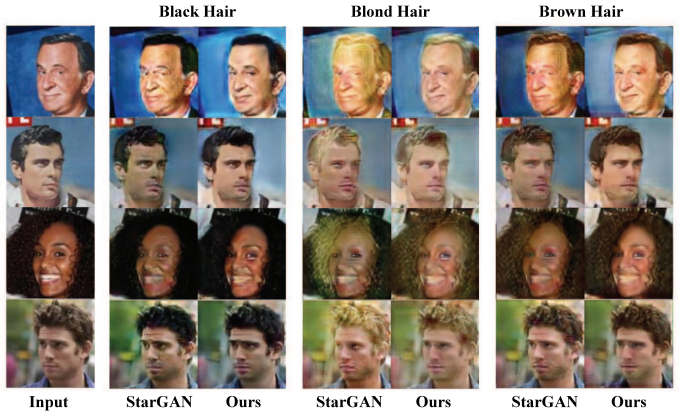

Fig. 9. The correctly classified images generated by $StarGAN$, and wrongly classified images generated by the proposed $StarGAN + IFS + GO$ for CelebA.

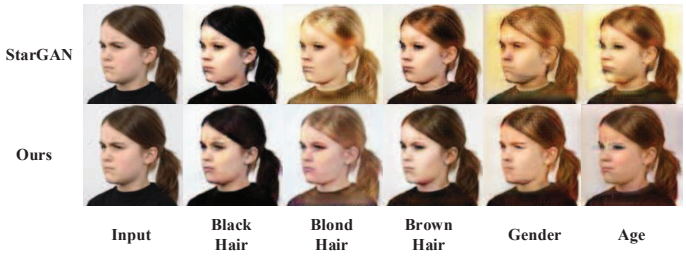

Fig. 10. The images generated by $StarGAN$ and the proposed GAN when CelebA and RaFD are used for training and testing, respectively.

the effectiveness of the proposed two strategies for semantic attribute disentanglement.

D. Generalization Performances

To study the generalization performance of the proposed algorithm, the CelebA and RaFD datasets are used for training and testing, respectively. The example images generated by $StarGAN$ and the proposed GAN for RaFD (training using CelebA) are presented in Fig. 10. Table IV presents the classification accuracies, IS and FID scores for this cross database experiment.

Fig. 10 reveals that the proposed GAN, i.e. $StarGAN + IFS + GO$, achieves more genuine performance than $StarGAN$, while yielding less abnormal regions on the nose and forehead regions, which reveals that the learned network with group-wise orthogonalization can be well generalized to other databases due to the reduction of the correlation among different semantic attributes.

Table IV further justifies the effectiveness of the proposed algorithm. Better performances with the proposed GAN are achieved on most of the attribute translations, in terms of the classification accuracy, IS and FID scores. Especially for the recognition accuracy of the translated black color attribute, an improvement of 26.67% is achieved for images generated by the proposed GAN with comparing to $StarGAN$.

TABLE IV
THE METRICS OF ACCURACY (ACC. %), IS AND FID OF $StarGAN$, $AttGAN$ AND OURS ($GAN_x + IFS + GO$) WHEN CELEBA AND RAFD ARE USED FOR TRAINING AND TESTING, RESPECTIVELY.

Metric	Method	Black	Blond	Brown	Gender	Age
Acc. ↑	$StarGAN$	53.75	**62.50**	57.29	70.63	66.67
	Ours	**80.42**	50.63	**60.63**	**74.79**	**77.50**
	$AttGAN$	56.46	**15.83**	43.33	41.04	50.42
	Ours	**59.17**	11.46	**49.58**	**44.38**	**74.38**
IS ↑	$StarGAN$	1.192	1.083	1.003	1.126	1.067
	Ours	**1.211**	**1.087**	**1.024**	**1.176**	**1.078**
	$AttGAN$	1.238	1.050	1.180	1.028	1.069
	Ours	**1.276**	**1.052**	**1.183**	**1.050**	**1.082**
FID ↓	$StarGAN$	138.74	169.35	150.26	176.01	188.59
	Ours	**136.85**	**163.49**	**142.78**	**161.10**	**172.46**
	$AttGAN$	**138.97**	191.56	**152.34**	184.23	171.03
	Ours	139.43	**184.73**	154.79	**170.23**	**156.69**

↑ means larger numbers are preferred, ↓ means opposite.

IV. CONCLUSIONS

In order to reduce semantic correlation caused by the feature map interaction to improve network generalization ability and allow independent editing of face attribute, we propose the semantic disentanglements of group-wise orthogonalization and intersection feature suppression in the generative networks of $StarGAN$ and $AttGAN$ for facial attribute translation. With the proposed semantic attribute disentanglement, the proposed GAN can synthesize much more genuine images with significantly less abnormality. Numerical results of classification accuracies, inception score and FID score further justify the effectiveness of the proposed GAN.

ACKNOWLEDGMENT

The work was supported by Natural Science Foundation of China under grants no. 61602315, 61672357 and U1713214, the Science and Technology Project of Guangdong Province under grant no. 2020A1515010707 and 2018A050501014, the Science and Technology Innovation Commission of Shenzhen under grant no. JCYJ20190808165203670.

REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NIPS*, pp. 2672–2680. 2014.
[2] Shiow Shung Yang and Ching-Shiow Tseng, "An orthogonal neural network for function approximation," *IEEE Trans Syst Man Cybern B Cybern*, vol. 26, no. 5, pp. 779–785, 1996.
[3] Chen Xi, Duan Yan, Rein Houthooft, John Schulman, and Pieter Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *NIPS*, pp. 2172–2180. 2016.
[4] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza, "Disentangling factors of variation for facial expression recognition," in *ECCV*, 2012, pp. 808–822.
[5] Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger, "Spectralnet: Spectral clustering using deep neural networks," in *ICLR*, 2018.
[6] Weishui Wan, Shingo Mabu, Kaoru Shimada, Kotaro Hirasawa, and Jinglu Hu, "Enhancing the generalization ability of neural networks through controlling the hidden layers," *Appl Soft Comput*, vol. 9, no. 1, pp. 404–414, 2009.
[7] Yongjun Hong, Uiwon Hwang, Jaeyoon Yoo, and Sungroh Yoon, "How generative adversarial networks and their variants work: An overview," *ACM Comput. Surv.*, vol. 52, no. 1, 2019.

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on August 13,2021 at 08:59:02 UTC from IEEE Xplore. Restrictions apply.

[8] Jie Liang, Jufeng Yang, Hsin-Ying Lee, Kai Wang, and Ming-Hsuan Yang, "Sub-gan: An unsupervised generative model via subspaces," in *ECCV*, 2018, pp. 698–714.
[9] Konstantin Klemmer, Adriano Koshiyama, and Sebastian Flennerhag, "Augmenting correlation structures in spatial data using deep generative models," *arXiv:1905.09796*, 2019.
[10] Yuxin Wu and Kaiming He, "Group normalization," in *ECCV*, 2018, pp. 3–19.
[11] Yunpeng Chen, Xiaojie Jin, Jiashi Feng, and Shuicheng Yan, "Training group orthogonal neural networks with privileged information," in *IJCAI*, 2017.
[12] Dong Wang, Lei Zhou, Xueni Zhang, Xiao Bai, and Jun Zhou, "Exploring linear relationship in feature map subspace for convnets compression," *arXiv preprint arXiv:1803.05729*, 2018.
[13] Sudipto Mukherjee, Himanshu Asnani, Eugene Lin, and Sreeram Kannan, "Clustergan: Latent space clustering in generative adversarial networks," in *AAAI*, 2019, vol. 33, pp. 4610–4617.
[14] Hadi Kazemi, Seyed Mehdi Iranmanesh, and Nasser Nasrabadi, "Style and content disentanglement in generative adversarial networks," in *WACV*, 2019, pp. 848–856.
[15] Luan Tran, Xi Yin, and Xiaoming Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *CVPR*, 2017, pp. 1415–1424.
[16] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos, "Deforming autoencoders: Unsupervised disentangling of shape and appearance," in *ECCV*, 2018, pp. 650–665.
[17] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou, "Interpreting the latent space of gans for semantic face editing," *arXiv:1907.10786*, 2019.
[18] Yunhe Wang, Chang Xu, Chao Xu, and Dacheng Tao, "Beyond filters: Compact feature map for portable deep model," in *ICML*, 2017, pp. 3703–3711.
[19] Yihui He, Xiangyu Zhang, and Jian Sun, "Channel pruning for accelerating very deep neural networks," in *ICCV*, 2017, pp. 1389–1397.
[20] Babajide O Ayinde and Jacek M Zurada, "Building efficient convnets using redundant feature pruning," in *ICLRW*, 2018.
[21] Babajide O Ayinde, Tamer Inanc, and Jacek M Zurada, "Redundant feature pruning for accelerated inference in deep neural networks," *Neural Networks*, vol. 118, pp. 148–158, 2019.
[22] Yi Tian, Zhiwei Wen, Weicheng Xie, Xi Zhang, Linlin Shen, and Jinming Duan, "Outlier-suppressed triplet loss with adaptive class-aware margins for facial expression recognition," in *ICIP*, 2019, pp. 46–50.
[23] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler, "Efficient object localization using convolutional networks," in *CVPR*, 2015, pp. 648–656.
[24] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018, pp. 8789–8797.
[25] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE Trans Image Process*, 2019.
[26] Hervé Abdi and Lynne J Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.
[27] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa, "L2-constrained softmax loss for discriminative face verification," *arXiv:1703.09507*, 2017.
[28] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao, "frame attention networks for facial expression recognition in videos," in *ICIP*, 2019, pp. 3866–3870.
[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015, pp. 3730–3738.
[30] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, Skyler T Hawk, and AD Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.
[31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved techniques for training gans," in *NIPS*, 2016, pp. 2234–2242.
[32] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NIPS*, 2017, pp. 6626–6637.
[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on August 13,2021 at 08:59:02 UTC from IEEE Xplore. Restrictions apply.