Learning Controllable Face Generator from Disjoint Datasets*

Jing Li[0000−0002−1384−7716], Yongkang Wong[0000−0002−1239−4428], and Terence Sim[0000−0002−0198−094X]

School of Computing, National University of Singapore
{lijing, wongyk, tsim}@comp.nus.edu.sg

Abstract. Recently, GANs have become popular for synthesizing photorealistic facial images with desired facial attributes. However, crucial to the success of such networks is the availability of large-scale datasets that are fully-attributed, *i.e.,* datasets in which the Cartesian product of all attribute values is present, as otherwise the learning becomes skewed. Such fully-attributed datasets are impractically expensive to collect. Many existing datasets are only partially-attributed, and do not have any subjects in common. It thus becomes important to be able to jointly learn from such datasets. In this paper, we propose a GAN-based facial image generator that can be trained on partially-attributed disjoint datasets. The key idea is to use a smaller, fully-attributed dataset to bridge the learning. Our generator (i) provides independent control of mutiple attributes, and (ii) renders photorealistic facial images with target attributes.

Keywords: Face generator · Disentanglement · Disjoint-learning.

1 Introduction

Research in facial image synthesis has spanned at least two decades in both computer graphics and computer vision. Much research employs machine learning approaches, and thus central to their success has been the collection and sharing of diverse facial image datasets that exhibit attribute variations, *i.e.,* images of subjects captured under different illumination conditions, head pose, facial expressions, *etc.*

However, many datasets (*e.g.,* CelebA [12], Multi-PIE [5]) exhibits a trade-off between subject diversity and image variations. Datasets acquired under laboratory settings typically contain a limited number ($\sim$ hundreds) of subjects, each captured under many different controlled imaging conditions; while datasets collected "in-the-wild" tend to consist of many ($\sim$ tens of thousands) subjects, each captured under only a few unknown and uncontrolled imaging conditions. For machine learning purposes, the ideal training dataset should contain a large number of subjects, each imaged under many different conditions. Moreover, all these attributes should be present in all possible combinations, as otherwise the learning would be skewed. Alas, collecting such fully-attributed dataset, where the full Cartesian product of all attribute values is present,

* This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its Strategic Capability Research Centres Funding Initiative.

is impractically expensive. Thus, it becomes important to learn from partially-attributed disjoint datasets, where each dataset exhibits variations in only one or a few attributes.

In fact, this joint-learning ability becomes *crucial* for face synthesis methods that employ deep networks, because such networks require large datasets for effective training. Recently, Generative Adversarial Network(GAN) [4] has been adopted to manipulate multiple attribute for facial images [2, 18].Unfortunately, these generators are hampered by their inability to learn from partially-attributed disjoint datasets.

In this paper, we propose a GAN-based facial image generator that can be learned from multiple partially-attributed disjoint datasets. The key idea is to use a smaller yet fully-attributed dataset to bridge the learning. Our proposed generator manipulates attributes for a facial image by extracting a disentangled attribute feature vector and by performing transformations on it. We demonstrate our method by learning to disentangle illumination, pose, and subject identity, and by rendering novel faces not seen in training under all variations of the said attributes. Our contributions are: **(1)** A novel GAN-based facial image generator with explicit control over attributes, thereby permitting the synthesis of arbitrary combinations of said attributes;**(2)** A learning method that learns from multiple partially-attributed disjoint datasets, thereby greatly increasing the learning capability of GAN-based generators. As a side benefit, our network also permits accurate analyses of facial image attributes.

2 Related Work

Face synthesis is widely studied in the literature. Compared with conventional methods [1, 6, 17] using handcraft feature models to capture attributes, deep learning methods [9, 18, 20] have been favored for image generation by taking advantage of vast real-life face images. For instance, Yang *et al.* [20] proposed a recurrent convolutional autoencoder network to rotate faces in images. Kulkarni *et al.* [9] presented a deep convolutional inverse graphics network, which learns disentangled pose, illumination, and identity features and allows for manipulation of multiple attributes.

Among deep learning methods, GAN-based models [4] has been favored for generating photorealistic images by training a generator against a discriminator. There are several works addressing multi-attribute manipulation for facial images. Li *et al.* [11] proposed a two-stage method for transfering attributes while preserving identity for a facial image. The method generates photorealistic face images but requires a set of reference facial images with target attribute for each input image. TD-GAN [18] learned to disentangle face attributes by integrating face generator with a tag mapping network to explore the consistency between images and their tags. Although this method is able to directly generate images from tags of target attributes, it requires for a set of training images with almost Cartesian product of all attribute variations. Recently StarGAN [2] allows simultaneous training of multiple datasets with different domains within a single network. It takes both image and domain information and learns to flexibly translate the image into corresponding domain. StarGAN is shown to perform well on synthesizing facial expressions of images from CelebA [12] dataset using features learned from RaFD [10] dataset. Although expression label of CelebA dataset are not explicitly used during training, there exist various facial expressions. Thus CelebA and RaFD are not

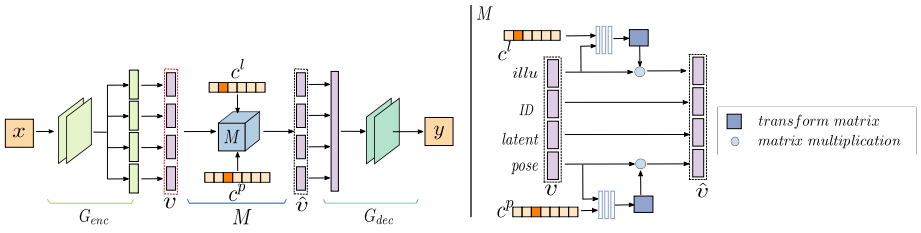

Fig. 1. Structure of proposed controllable multi-attribute face generator network. The face generator is composed of an encoder G_{enc}, an attribute manipulator network M, and a decoder G_{dec}. G_{enc} encodes a facial image into illumination, pose, ID and latent subspaces.

completely disjoint with respect to facial expression. Furthermore, manipulation of facial attributes such as expressions and hair color, involves only local changes of images. It is unclear how effective it is to the manipulation of global facial attributes (*e.g.,* pose).

All the above methods suffer from a serious drawback: they require a training dataset that is fully-attributed. This makes them impractical since such datasets are expensive to collect. It would be more practical to be able to learn from partially-attributed disjoint datasets, and still retain the ability to control rendering attributes, while generating realistic images. This paper proposes such a method: the key idea is to use a small bridging dataset that mediates between the disjoint datasets.

3 Proposed Face Generator Framework

We propose a face generator network (Figure 1) that learns disentangled and discriminative embedding for different attributes, including illumination, pose, ID and latent information useful for image rendering. Within this embedding, an attribute manipulation network, shown on the right panel of Figure 1, is devised to modify illumination and pose feature by learning a transform matrix from two one-hot vectors (*i.e.,* c^l and c^p) that indicates target illumination and pose.

We train the proposed network for two disjoint facial image datasets: $\mathcal{A}$ with frontal pose but variable illuminations and $\mathcal{C}$ in reverse. To address the attribute disentanglement and rendering of images with attributes absent in both $\mathcal{A}$ and $\mathcal{C}$ (*e.g.,* facial images with 45° illumination and 45° pose), we employ a small fully-attributed bridging dataset $\mathcal{B}$, which can be easily collected in practice as it requires only a few subjects. The training method (shown in Figure 2) works by applying two-stream attribute manipulations on an input facial image: one is to manipulate illumination to target c^l and the other is to manipulate pose to target c^p. We train the proposed model with following objectives.

Attribute Disentanglement and Discriminability. These are two desirable properties for attribute subspace, because disentanglement allows for changing an attribute for a facial image without tempering others, and discriminative attribute subspace makes it easy for attribute manipulation operation and attribute interpolation.

To achieve this objective, we adopt triplet loss [14](denoted as L_{tri}) in each of illumination, pose and ID subspaces learned by encoder G_{enc}[1]. The intrinsic idea of

[1] For convenience, we normalized each feature vector in embedding subspaces in network.

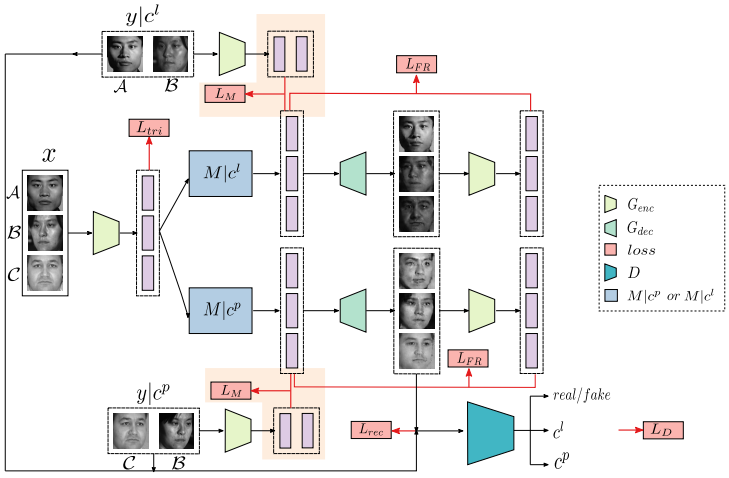

Fig. 2. Overview of proposed method. For each image x, the model changes it separately to illumination c^l and pose c^p, resulting in two images.

triplet loss is to minimize intra-class distance and maximize inter-class distance. While triplet loss is designed for training discriminative feature subspace, it also achieves attribute disentanglement. Take illumination as an example, pose and identity variations are untangled from illumination attribute subspace, because facial images with same illumination but different poses and identities are enforced to be close in the subspace by triplet loss objective. Similarly, illumination and identity are untangled in pose subspace, and illumination and pose are untangled in ID subspace.

Attribute Manipulation. This is performed by transforming a normalized attribute feature vector from one class to target class. We conduct two separate manipulations on illumination and pose, denoted as M^{illu} and M^{pose} for training images. We pair each image $\boldsymbol{x}$ with a target image $\boldsymbol{y}$ conditioned on M^{illu}/M^{pose}, and enforce the transformed attribute feature vector of $\boldsymbol{x}$ to be same with that of $\boldsymbol{y}$. Nevertheless, there is no target image for M^{pose} of $\mathcal{A}$ images and M^{illu} of $\mathcal{C}$ images. We hence select a pseudo target image from the other dataset. We denote image pairs as *homogeneous pair* and *heterogeneous pair* as Figure 3. Homogeneous pairs come from a same dataset and differs in only one particular attribute. We propose a loss function for homogeneous pairs as:

$$L_{h1}(\mathcal{S}, attr) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\mathcal{S}}\left[\, \|\boldsymbol{v}_{\boldsymbol{x}}^{\overline{attr}} - \boldsymbol{v}_{\boldsymbol{y}}^{\overline{attr}}\|_2^2 + \|\boldsymbol{v}_{\boldsymbol{x}}^{ID} - \boldsymbol{v}_{\boldsymbol{y}}^{ID}\|_2^2 + \|M^{attr}(\boldsymbol{v}_{\boldsymbol{x}}^{attr}, c^{attr}) - \boldsymbol{v}_{\boldsymbol{y}}^{attr}\|_2^2 \,\right] \quad (1)$$

where $\mathcal{S} \in \{\mathcal{A}, \mathcal{B}, \mathcal{C}\}$, $attr \in \{illu, pose\}$ and $\overline{attr} = \{illu, pose\} - attr$. We sum up the homogeneous loss as:$L_{ho} = L_{h1}(\mathcal{A}, illu) + L_{h1}(\mathcal{C}, pose) + L_{h1}(\mathcal{B}, illu) + L_{h1}(\mathcal{B}, pose)$. In homogeneous pairs, images come from two partially-attributed datasets with different attributes. Therefore, only the distance between the transformed attribute feature vector and its pseudo target attribute feature is considered in heterogeneous loss:

$$L_{he} = \mathbb{E}\left[\, \|M^{pose}(\boldsymbol{v}_{\boldsymbol{x}_{\mathcal{A}}}^{pose}, c^p) - \boldsymbol{v}_{\boldsymbol{y}_{\mathcal{C}}}^{pose}\|_2^2 + \|M^{illu}(\boldsymbol{v}_{\boldsymbol{x}_{\mathcal{C}}}^{illu}, c^l) - \boldsymbol{v}_{\boldsymbol{y}_{\mathcal{A}}}^{illu}\|_2^2 \,\right] \quad (2)$$

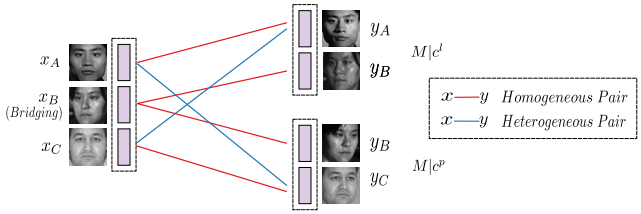

Fig. 3. Attribute transform pairs. Homogeneous images are from same dataset with same identity, and heterogeneous images are from two disjoint datasets with different identity. Each pair demonstrates variation in either illumination or pose.

Photorealistic Image Generation. This is achieved by regression loss and GAN loss. Firstly, since the face generator follows autoencoder architecture, we use L2 loss as L_{rec} for homogeneous pairs. In addition, we train face generator against a discriminator D, and employ GAN loss for realistic facial image generation by:

$$L_{GAN} = \mathbb{E}_{\boldsymbol{x}\sim r} \log D(\boldsymbol{x}) + \mathbb{E}_{\boldsymbol{x}\sim g} \log(1 - D(\boldsymbol{x})) \tag{3}$$

In order to encourage face generator generate images with correct target illumination and pose, we apply two 1-of-2N classifiers [15], C, on top of feature layers of discriminator to predict illumination and pose of the generated images. 1-of-2N classifier is designed with the purpose of classifying a real image $\boldsymbol{x}_r$ to first N labels and a generated image $\boldsymbol{x}_g$ to the last N labels, where N is the number of attribute classes. By considering a same attribute class of real and generated images as different label, adversarial training pushes real and generated domains as close to each other as possible, thus preserving attributes of generated images. Cross-entropy loss is used to train C:

$$L_{C_{attr}} = \textstyle\sum_j -\{[c_r^{attr}\ \mathbf{0}]\}_j \log(\{C_{attr}(\boldsymbol{x}_r)\}_j) + \sum_j -\{[\mathbf{0}\ c_g^{attr}]\}_j \log(\{C_{attr}(\boldsymbol{x}_g)\}_j) \tag{4}$$

where $\mathbf{0}$ is a N-dim zero vector, $j \in [1, 2N]$ is the j-th index of attribute class.

Identity Preservation. In proposed method, discriminator D is employed to enforce the distribution of generated images to be similar to that of real images. However, due to that $\mathcal{A}$ and $\mathcal{C}$ are partially-attributed, the real images with non-frontal illumination and non-frontal pose for D only comes from $\mathcal{B}$. This would result in that the generated images $M^{pose}(\boldsymbol{x}_\mathcal{A})$ and $M^{illu}(\boldsymbol{x}_\mathcal{C})$ appear to be like faces of $\mathcal{B}$ with same attribute conditions. To address this problem, we propose a feature reconstruction loss to preserve the facial identity of generated images. Specifically, we extract the semantic feature of each generated image by the encoder G_{enc}, and enforce the semantic feature to be same as the one that is used for generating the current image. See Figure 2 for illustration. The feature reconstruction loss is presented as:

$$L_{FR} = \mathbb{E}_{\boldsymbol{v}\sim G_{enc}(\boldsymbol{x})} \|\boldsymbol{v} - G_{enc}(G_{dec}(\boldsymbol{v}))\|_2^2 \tag{5}$$

Implementation Details. We set (64, 64, 256, 64) for dimensions of illumination, pose, ID and latent features in proposed networks. We trained our face generator by:

$$L = L_{rec} + \lambda_1(L_{tri}^{illu} + L_{tri}^{pose} + L_{tri}^{ID}) + \lambda_2(L_{ho} + L_{he}) + \lambda_3 L_{FR} + \lambda_4(L_{GAN} + L_{C_{illu}} + L_{C_{pose}}) \tag{6}$$

Table 1. Details of experimental datasets. Note that we label 67° illumination of CMU-PIE to be same with 90° illumination of Multi-PIE and CAS-PEAL, as they are visually similar in images.

Dataset	Subjects	Images	Illumination	Pose	light
CMU-PIE [16]	68	2380	$0^\circ, \pm 30^\circ, \pm 67^\circ$	$0^\circ, \pm 15^\circ, \pm 30^\circ, \pm 45^\circ$	off
Multi-PIE [5]	336	11760	$0^\circ, \pm 45^\circ, \pm 90^\circ$	$0^\circ, \pm 15^\circ, \pm 30^\circ, \pm 45^\circ$	on
CAS-PEAL [3]	233	920	$0^\circ, \pm 45^\circ, \pm 90^\circ$	0°	off

Table 2. Experimental configurations. • denotes single attribute face dataset and $*$ denotes the small scale bridging face dataset. Note that subjects of each subset in the table is non-overlapped.

Scenario	Training Data	Data Source
I (Ideal)	Fully-attribute dataset with 200 subjects	Multi-PIE
II (w/o bridge)	• Pose attribute subset with 100 subjects	Multi-PIE
	• Illumination attribute subset with 100 subjects	Multi-PIE
III (w/ bridge)	• Pose attribute subset with 100 subjects	Multi-PIE
	• Illumination attribute subset with 100 subjects	Multi-PIE
	$*$ Bridging dataset with 20 subjects	Multi-PIE
IV (cross dataset)	• Pose attribute subset with 200 subjects	Multi-PIE
	• Illumination attribute subset with 200 subjects	CAS-PEAL
	$*$ Bridging dataset with 20 subjects	CMU-PIE

where λ_1, λ_2, λ_3 and λ_4 are set to 0.01, 0.5, 0.5, and 0.001. We set triplet margins for illumination, pose and identity features to be 1.8, 1.8, and 0.7. Adam optimizer is used to train the G and M with learning rate at 0.001 and decay rate at 0.01, and D with learning rate at 0.001 and decay rate at 0.01.

4 Experiments

Dataset and Preprocessing. We evaluated the proposed generator on CMU-PIE [16], Multi-PIE [5] and CAS-PEAL [3] datasets (see Table 1). All images were aligned using chin and forehead position detected by Dlibrary [8] and cropped to 128×128 pixels.

Experimental Configuration. We considered four scenarios to investigate how face generator performs with different training dataset configurations (Table 2). Scenario I is the ideal training configuration where each subject has multiple images captured under all factors (*i.e.,* controlled poses and illumination conditions). It benchmarks the best performance that can be achieved by a face generator. In Scenario II, the training was conducted using two face datasets with single attribute variation. This scenario stresses testing the proposed model under extreme learning condition. In Scenario III, we added on top of Scenario II a bridging dataset with small number of subjects. The newly added bridging dataset provided *guidance information* during model training stage. In Scenario IV, we designed a configuration that mimics realistic scenario, where partially-attributed datasets come from different data sources.

Evaluation. We evaluated the learned face generator with two aspects, namely feature subpaces and image quality. The feature subpaces were evaluated by performing a series of classification tasks on attribute feature vectors, and the image quality was evaluated by Fréchet Inception Distance (FID) [7], a measure of similarity between the generated images and real images.

Table 3. Classification accuracy on feature subspaces for Multi-PIE experiments.

		Illumination	Pose	ID
Scenario I (Ideal)	Illu Subspace	**100.00**%	14.68%	0.20%
	Pose Subspace	20.39%	**100.00**%	0.0%
	ID Subspace	29.09%	20.56%	**100.00**%
Scenario II (w/o bridge)	Illu Subspace	66.36%	35.76%	0.86%
	Pose Subspace	43.03%	80.04%	0.51%
	ID Subspace	64.03%	50.87%	83.92%
Scenario III (w/ bridge)	Illu Subspace	99.61%	15.24%	0.51%
	Pose Subspace	24.11%	93.72%	0.15%
	ID Subspace	27.84%	19.22%	99.79%

Table 4. Image quality evaluation for Multi-PIE experiments.

	FID	Illumination	Pose	ID
Scenario I (Ideal)	19.59	99.15%	95.59%	91.50%
Scenario II(w/o bridge)	139.81	36.69%	70.69%	65.65%
Scenario III(w/ bridge)	37.79	98.15%	95.04%	86.07%

4.1 Experiments on Multi-PIE

We firstly investigated the importance of a fully-attribute training dataset to the learning of a multi-attribute face generator by comparing our method in Scenario I, II and III.

Feature Subspaces. We evaluated the discriminability and disentanglement of semantic attribute subspaces by performing classification tasks on the testing images of 116 subjects in Multi-PIE. For illumination and pose classification, we used images of 50 subjects as galleries and images of the remaining 66 subjects as probes. For face verification, we randomly selected 17 images for each of 116 subjects as galleries and used the remaining 18 images as probes. We employed KNN method using Euclidean Distance metric for each task.

We showed the classification results in Table 3. As can be seen, the baseline Scenario I achieved the best performance, 100% accuracy, for all three classification tasks. However, the performance quickly drops in Scenario II, where much lower accuracy rates (66.36%, 80.04%, 83.92%) were attained for illumination, pose classifications and face verification in corresponding subspace. Moreover, it can be noted that substantial results were achieved in unassociated subspaces for illumination and pose classification tasks. This is because that the training data sets were partially-attributed, hence the learned model cannot disentangle facial images with non-frontal illumination and non-frontal pose. With a small fully-attributed bridging dataset in Scenario III, we attained classification performance that approaches ideal Scenario I.

Image Quality. We evaluated quality of generated images by computing FID scores between synthesized images and real images of testing Multi-PIE subjects. We also evaluated attribute preservation of generated images by performing illumination and pose classifications as well as face verification on generated images of testing subjects. We used the testing images as galleries and the generated images as probes. We conducted the experiments by using a pretrained face network [19] as feature extractor, PCA for reducing feature dimension and then SVM classfication method. The results are shown in Table 4, in which Scenario I achieves the best performance for all evalu-

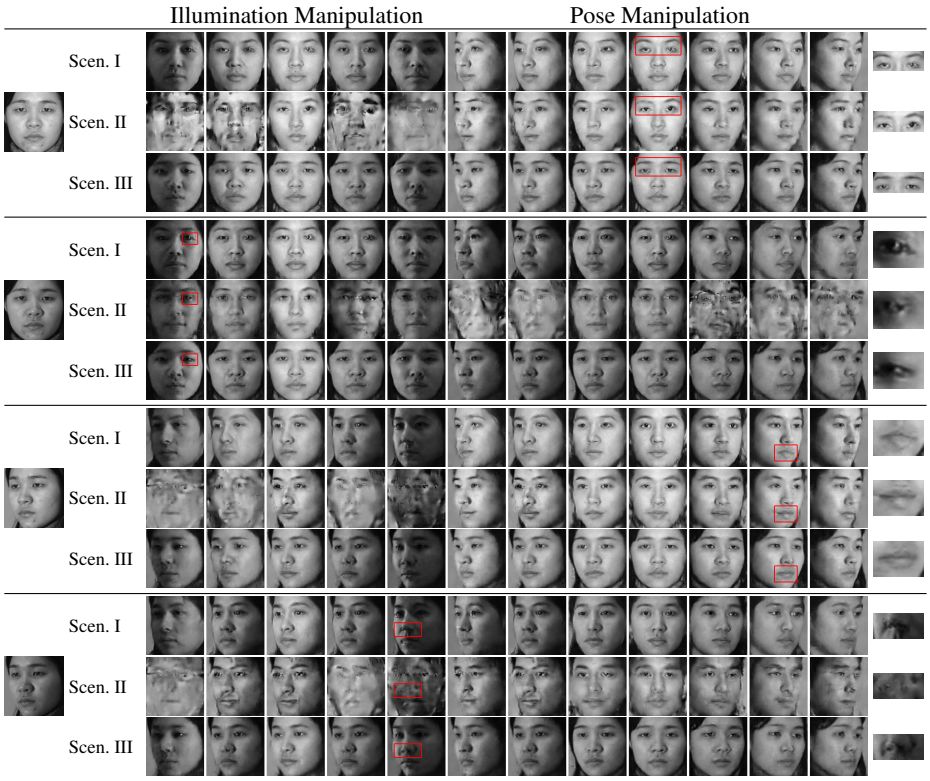

Fig. 4. Comparison of generated images for models in Scenario I, Scenario II and Scenario III. Each face was changed in illumination and pose, respectively. It is clear that the use of the bridging dataset (Scenario III) improves image quality.

ation on generated images with smallest FID score and largest classification accuracy rates. In scenario II, the quality of generated images drops to 139.81. This is mainly because the model learned was unable to generate photorealistic images with unseen illumination and pose. By using a small bridging dataset, our method was able to synthesize images with target attributes and preserved facial identity, where FID score was reduced to 37.79, and classification performance was improved by 61.46%, 24.35%, 20.42% for illumination, pose and identity, respectively.

We illustrated some example images in Figure 4. As can be seen, model II failed to generate convincing results, especially images with illumination changed. This is consistent with the low illumination classification accuracy using feature vectors(66.36%) and synthesized images(36.69%). It also fell short of changing poses for facial images with non-frontal illumination absent during training (face 2, 4). On the other hand, model I and III succeeded to generate photorealistic facial images, even if testing images were at random pose and illumination. Generally, model I achieved better holistic image quality, while model III preserved better local details, such as eye regions in face 1 and 2. One possible reason is that model I learned more variations for each attribute condition and failed preserve local details while trying to capture global structure.

Table 5. Classification accuracy on feature spaces for cross dataset experiments.

	Illumination	Pose	ID
TD-GAN [18]	96.88%	31.34%	7.05%
Our method	**99.72%**	**100.00%**	**99.78%**

Table 6. Image quality evaluation for cross datasets experiments.

Method	FID			Illumination	Pose	ID
	Multi-PIE	CAS-PEAL	CMU-PIE			
TD-GAN [18]	310.49	377.58	329.49	94.25%	59.34%	54.86%
StarGAN [2]	118.41	172.31	142.60	**95.09%**	97.55%	**80.72%**
Ours	34.33	139.02	55.35	94.83%	**98.51%**	69.97%

4.2 Cross Dataset Experiments

We compared our method with existing methods, namely TD-GAN [18] and StarGAN [2], in Scenario IV, where each subset of training data came from different datasets.
Feature Subspaces. Both our method and TD-GAN learn disentangled attribute feature subspaces for input images. In the same line, we compared their feature subspaces by performing a series of classification tasks. We used images of 116 subjects from Multi-PIE, 33 subjects from CAS-PEAL, and 48 subjects from CMU-PIE as testing data. For illumination and pose classification, we use CMU-PIE images as galleries and Multi-PIE as well as CAS-PEAL images as probes. For face verification, we split images of each subject into two equal subsets, one for training and the other for testing.

The results were presented in Table 5. As can be seen, our method attained higher accuracy rates in all tasks than TD-GAN. This is because our method concerns both inter-class and intra-class distance for attribute feature vectors, whereas TD-GAN only considers inter-class distance by enforcing feature vectors of same class to be close.

We plotted the feature subspaces in Figure 5 using t-SNE [13]. It can be seen that TD-GAN network failed to disentangle face attributes, while our method performed well in attribute disentanglement. The main reason for TD-GAN's poor performance is that it employed L2 loss to enforce an attribute feature of an image to be close to that learned from its label. However, without considering inter-class distance, L2 loss leads to the trivial feature vectors wher their norm $\|\boldsymbol{v}\|_2^2$ equals to 0.

Image Quality. We evaluated quality of generated images by FID scores and attribute classifications performance on testing images of all datasets. Being reminded that CAS-PEAL and Multi-PIE images used here were partially-attributed, so we only compared generated partially-attributed images for these datasets with testing data to compute FID score. Also, we examined the attribute accuracy of generated images by performing illumination, pose classification and face identification tasks on galleries and probes as used during feature analysis metioned above.

We presented the results in Table 6. It was shown that our method generated more photorealistic facial images than other two methods, with the smallest FID scores for all datasets. In addition, our method was able to generate images with target attributes, with illumination and pose classification accuracy rates at 94.83% and 98.51%. In face verification, our method outperformed TD-GAN yet performed worse than StarGAN. This is because our method learns high-level attribute features in bottleneck layer, and

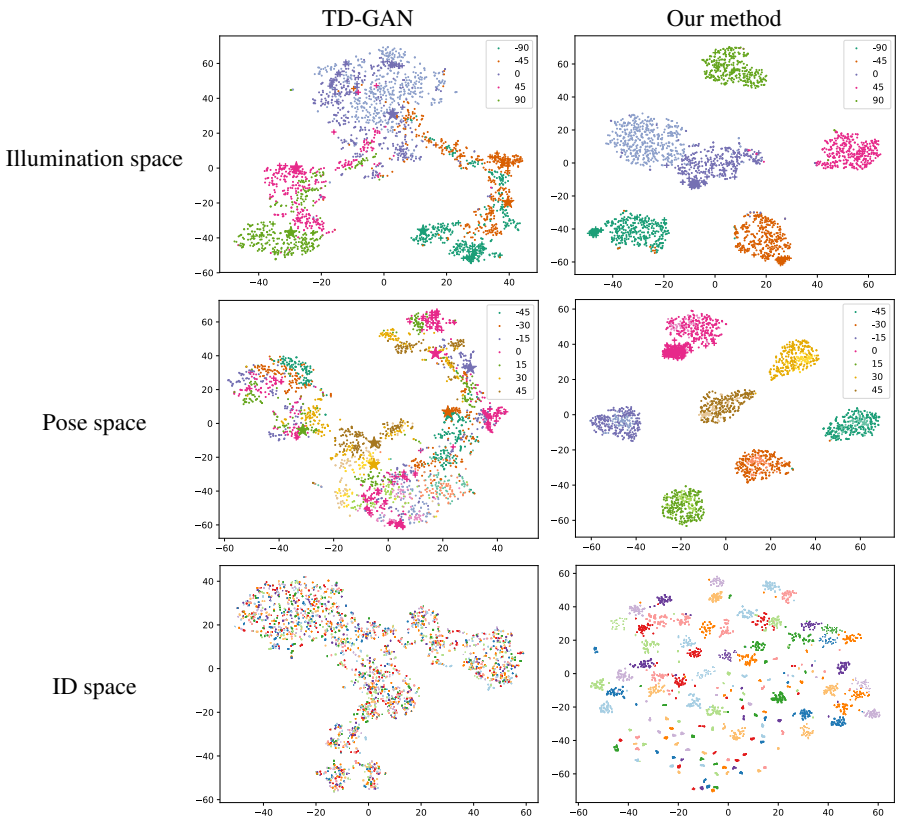

Fig. 5. (Best viewed in color) Visualization of feature subspaces of TD-GAN(LEFT) and our method(RIGHT). illumination and pose class center learned by tag network as $\star$ in the figures. We denote $\bullet$ for CMU-PIE images, $+$ for CAS-PEAL images and use the lighter color for Multi-PIE images. Our method achieves better separation of illumination, pose and identity attributes, leading to more accurate classification of said attributes.

therefore fails to capture high frequency image information. Whereas StarGAN directly learns the attribute translation in the network and can preserve low-level image feature.

The generated images are shown in Figure 6. As shown, TD-GAN failed to generate photorealistic facial images. StarGAN was able to preserve the identity of synthesized images as it directly learns the attribute translation method without learning semantic feature space. Yet StarGAN performed poorly in pose manipulation that involves spatial change in images (see the ghosting effects in pose manipulation results). In general, our method can generates images with reasonable image quality and preserves identity.

5 Conclusion

In this paper, we presented a face generator that learns from partially-attributed disjoint datasets, along with a smaller fully-attributed briding dataset. The proposed method al-

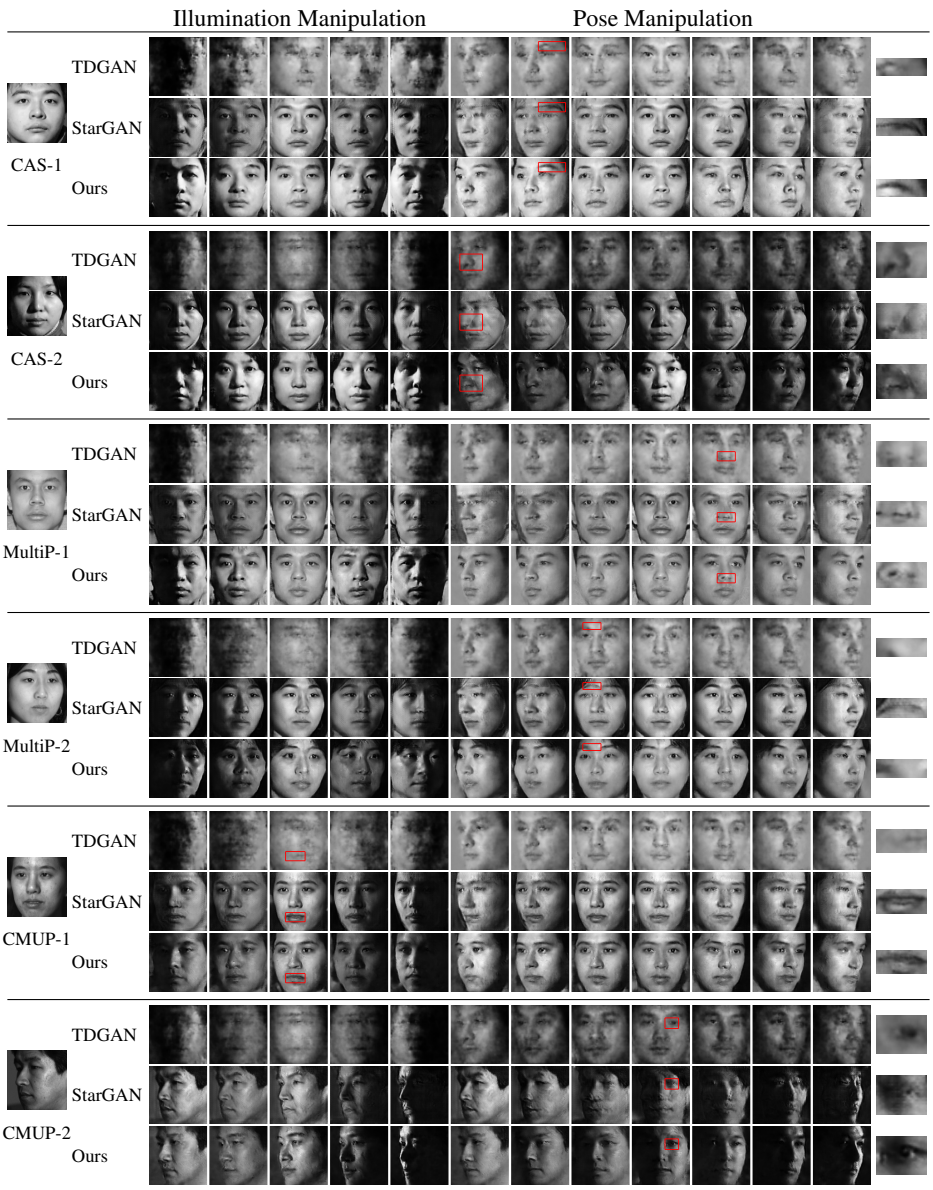

Fig. 6. Comparison of generated images for our method and TD-GAN, StarGAN. Each face was changed in illumination and pose, respectively. Our method renders images with better visual quality.

lows for explicit control over multiple attributes for a facial image by learning a disentangled and discriminative feature space. We conducted experiments under four training scenarios and showed that by using a small bridging dataset, the disentanglement and rendering of multiple attributes for partially-attributed disjoint datasets can be easily ad-

dressed. Compared with TD-GAN and StarGAN, our method renders superior images in both illumination and pose variations. As a side benefit, our network also achieves higher classification accuracy of the said attributes.

References

1. Almaddah, A., Vural, S., Mae, Y., Ohara, K., Arai, T.: Face Relighting using Discriminative 2D Spherical Spaces for Face Recognition. Mach. Vis. Appl. pp. 845–857 (2014)
2. Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., Choo, J.: StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In: CVPR 2018. pp. 8789–8797
3. Gao, W., Cao, B., Shan, S., Chen, X., Zhou, D., Zhang, X., Zhao, D.: The CAS-PEAL Large-Scale Chinese Face Database and Baseline Evaluations. IEEE Trans. Systems, Man, and Cybernetics, Part A pp. 149–161 (2008)
4. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: NIPS 2014. pp. 2672–2680
5. Gross, R., Matthews, I.A., Cohn, J.F., Kanade, T., Baker, S.: Multi-PIE. Image Vision Comput. **28**(5), 807–813 (2010)
6. Heo, J., Savvides, M.: 3-D Generic Elastic Models for Fast and Texture Preserving 2-D Novel Pose Synthesis. IEEE Trans. Information Forensics and Security **7**(2), 563–576 (2012)
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: NIPS 2017. pp. 6629–6640
8. King, D.E.: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research **10**, 1755–1758 (2009)
9. Kulkarni, T.D., Whitney, W.F., Kohli, P., Tenenbaum, J.B.: Deep Convolutional Inverse Graphics Network. In: NIPS 2015. pp. 2539–2547
10. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.: Presentation and validation of the radboud faces database. Cognition and emotion **24**(8), 1377–1388 (2010)
11. Li, M., Zuo, W., Zhang, D.: Convolutional Network for Attribute-driven and Identity-preserving Human Face Generation. CoRR **abs/1608.06434** (2016)
12. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV 2015. pp. 3730–3738
13. Maaten, L.v.d., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**(Nov), 2579–2605 (2008)
14. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: CVPR 2015. pp. 815–823
15. Shen, Y., Luo, P., Yan, J., Wang, X., Tang, X.: Faceid-gan: Learning a symmetry three-player GAN for identity-preserving face synthesis. In: CVPR 2018. pp. 821–830
16. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression (PIE) database. In: AFGR 2002. pp. 53–58
17. Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., Theobalt, C.: Real-time expression transfer for facial reenactment. ACM Trans. Graph. pp. 183:1–183:14 (2015)
18. Wang, C., Wang, C., Xu, C., Tao, D.: Tag Disentangled Generative Adversarial Networks for Object Image Re-rendering. In: IJCAI, 2017
19. Wu, X., He, R., Sun, Z., Tan, T.: A light CNN for deep face representation with noisy labels. IEEE Trans. Information Forensics and Security **13**(11), 2884–2896 (2018)
20. Yang, J., Reed, S.E., Yang, M., Lee, H.: Weakly-supervised Disentangling with Recurrent Transformations for 3D View Synthesis. In: NIPS 2015. pp. 1099–1107