Instance-Level Facial Attributes Transfer with Geometry-Aware Flow

Weidong Yin
University of British Columbia
wdyin@cs.ubc.ca

Ziwei Liu
Chinese University of Hong Kong
zwliu@ie.cuhk.edu.hk

Chen Change Loy
Nanyang Technological University
ccloy@ntu.edu.sg

Abstract

We address the problem of instance-level facial attribute transfer without paired training data, *e.g.,* faithfully transferring the exact mustache from a source face to a target face. This is a more challenging task than the conventional semantic-level attribute transfer, which only preserves the generic attribute style instead of instance-level traits. We propose the use of geometry-aware flow, which serves as a well-suited representation for modeling the transformation between instance-level facial attributes. Specifically, we leverage the facial landmarks as the geometric guidance to learn the differentiable flows automatically, despite of the large pose gap existed. Geometry-aware flow is able to warp the source face attribute into the target face context and generate a warp-and-blend result. To compensate for the potential appearance gap between source and target faces, we propose a hallucination sub-network that produces an appearance residual to further refine the warp-and-blend result. Finally, a cycle-consistency framework consisting of both attribute transfer module and attribute removal module is designed, so that abundant unpaired face images can be used as training data. Extensive evaluations validate the capability of our approach in transferring instance-level facial attributes faithfully across large pose and appearance gaps. Thanks to the flow representation, our approach can readily be applied to generate realistic details on high-resolution images[1].

Introduction

Modeling and manipulating facial attributes have been a long quest in computer vision (Liu et al. 2015; Nguyen et al. 2008; Huang et al. 2018; Loy, Luo, and Huang 2017). On the one hand, facial attributes are one of the most prominent visual traits we perceive in daily life, thus constituting an important visual element to understand. On the other hand, the ability to manipulate facial attributes can enable lots of useful real-world applications, such as targeted face editing (Shu et al. 2017; Shen and Liu 2017; Brock et al. 2016; Yeh et al. 2016).

Existing studies on facial attribute manipulation mainly focus on semantic-level attribute transfer (Perarnau et al. 2016; Lample et al. 2017; Choi et al. 2017; Gardner et al.

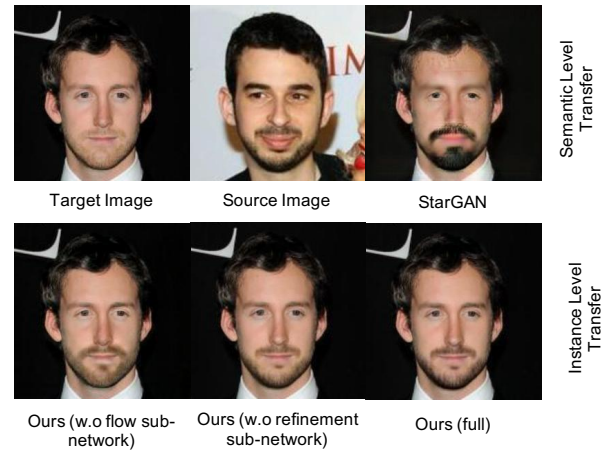[1]Project page: http://mmlab.ie.cuhk.edu.hk/projects/attribute-transfer/

Figure 1: (First row) Semantic-level facial attribute transfer, and (second row) instance-level facial attribute transfer. The former only models the generic attribute style while the latter additionally preserves sample-dependant traits.

2015), *i.e.,* making the target face possess certain attributes (*e.g.,* 'mustache') perceptually, as shown in Figure 1 (first row). There is, however, no guarantee that the transferred 'mustache' on the target face will look alike the one on the source face.

In this work, we address the problem of instance-level facial attribute transfer without paired training data. For example, given an unordered collection of images with and without mustache for training, our approach learns to faithfully transfer the exact mustache from the target face to the source face. This is a more challenging task than the conventional semantic-level attribute transfer. In particular, besides capturing the generic attribute style that are shared by all 'mustache' samples, instance-level attribute transfer requires the extraction and preservation of sample-dependant traits, *e.g.,*, the mustache from the source image and the lip of the target image, as shown in Figure 1 (second row).

The main contribution of our paper is a novel notion of ***geometry-aware flow***, which is well-suited as a representation to model the transformation between instance level attributes. Unlike the conventional differentiable flow (Jader-
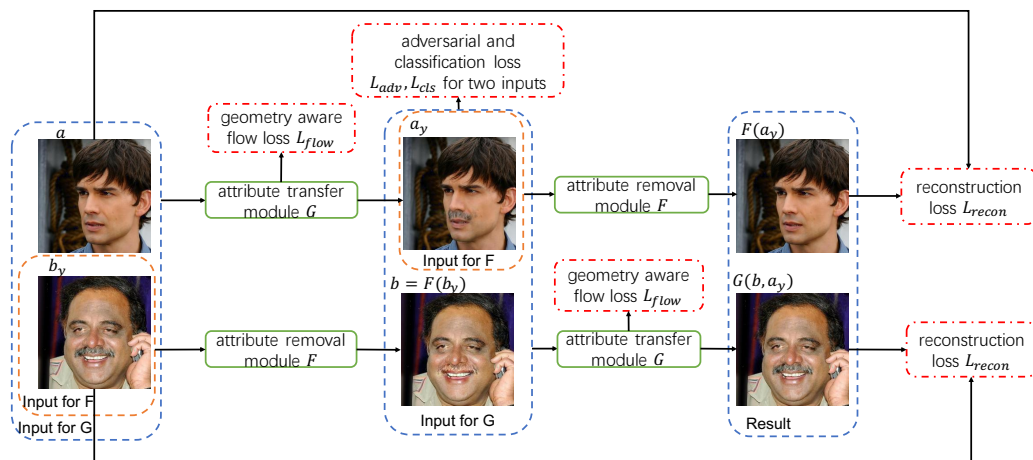
Figure 2: The pipeline of instance-level attribute transfer. Given a target image a and a source image b_y, our approach automatically learns an attribute transfer module G and an attribute removal module F. These two modules jointly operate in a cycle-consistency manner to learn from abundant unpaired data.

berg et al. 2015; Zhou et al. 2016), geometric flow is learned under the geometric guidance from facial landmarks. Consequently, the flow can cope robustly with the large displacement gap between facial poses of target and source images. In addition, since flow is invariant to scaling, the learned geometry-aware flow can be readily applied to high-resolution images and generate desired attributes with realistic warp-and-blend details.

To further enhance the transfer quality, we propose a refinement sub-network to rectify potential appearance gaps, *e.g.,* , skin color and lighting changes, which cannot be handled well with flow-based warping. This is achieved through generating an appearance residual that can be added to the previous warp-and-blend result. The importance of the refinement network can be observed in Figure 1 (second row). Without the sub-network, the transferred 'mustache' on the target face is susceptible to the surrounding skin color of the target face. Finally, a cycle-consistency framework consisting of an attribute transfer module and an attribute removal module is designed so that learning can be done on abundant unpaired facial images. Extensive evaluations on CelebA (Liu et al. 2015) and CelebA-HQ (Karras et al. 2017) datasets validate the effectiveness of our approach in transferring instance-level facial attributes faithfully across large pose and appearance gaps.

Related Work

Semantic-level Transfer. Many recent works have achieved impressive results in this direction. StarGAN (Choi et al. 2017) applies cycle consistency to preserve identity, and uses classification loss to transfer between different domains. The task of facial attribute transfer can also be viewed as a specific kind of image-to-image translation problem. UNIT (Liu, Breuel, and Kautz 2017) combines variational autoencoders (VAEs) (Kingma and Welling 2013) with Co-GAN (Liu and Tuzel 2016) to learn the joint distribution of images in two domains. These methods only tackle the task of facial attribute transfer at a semantic level and could not transfer specific attribute of different style at an instance level.

Instance-level Transfer. There exist some recent studies that discuss the possibility of transferring facial attributes at instance level. GeneGAN (Zhou et al. 2017) learns to transfer a desired attribute from a reference image by constructing disentangled attribute subspaces from weakly labeled data. ELEGANT (Xiao, Hong, and Ma 2018) exchanges attribute between two faces by exchanging latent codes of two faces. All these methods are fully parametric, which suffers from blurry results and cannot scale to high-resolution. Some other studies try to directly compose exemplar images. ST-GAN (Lin et al. 2018) uses a spatial transformer network to transform external objects into the correct position before composing them onto faces. Compositional GAN (Azadi et al. 2018) learns to model object compositions in a GAN framework. Different from these works, our approach does not require an object to be provided explicitly. Instead, we extract the desired facial attribute automatically from exemplar images to facilitate the faithfulness of instance-level attribute transfer.

Methodology

Let A be a domain of target images without certain attribute and B be a domain of source images with the specific kind of attribute we want to transfer, where no pairings exist between the two domains. Given $a \in A$ be a target image without attribute and $b_y \in B$ a source image with the specific kind of attribute we want to transfer, where in the case of people's faces are selected as $eyeglasses, mustache, goatee$. Our goal is to learn a model G that can transfer desired attributes from b_y onto a such that we obtain a_y after the transfer. Different from general attribute manipulation (Choi et al. 2017), our model transfers attribute at instance level. For the same target image, given different source images we can transfer the same at-
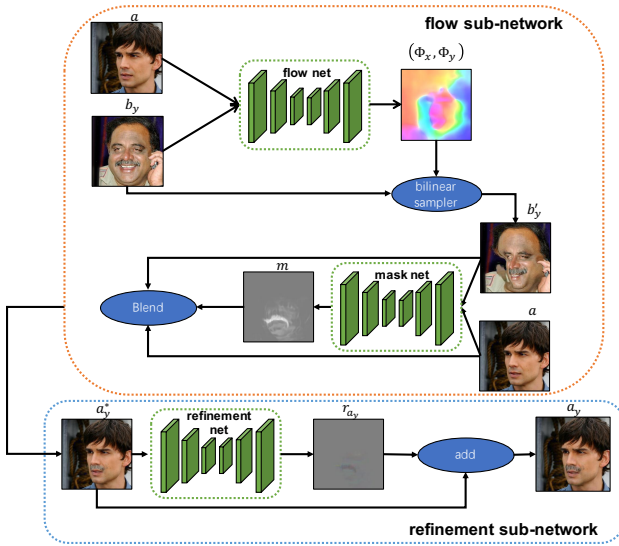
Figure 3: Network architecture of attribute transfer module. Given a source image and a target image, the flow sub-network learns to generate geometric-aware flow to warp the source image into the desired pose. An attention mask is then generated to blend the source and target images. Next, the refinement sub-network synthesizes an appearance residual to compensate for other differences in appearance, like skin colors or lighting conditions.

tribute of different styles onto the target image to generate diverse results.

Instance-level Attribute Transfer Network

As shown in Figure 2, two separate modules are introduced to achieve instance-level attribute transfer.

- Attribute transfer module G. Given an image of a person without some attribute a, and an image of a different person with desired attribute b_y, the attribute transfer module $G : A \times B \to B$ extracts the desired attribute from b_y and applies it to a maintaining its identity.

- Attribute removal module F. Given the same photo b_y, the attribute removal network $F : B \to A$ learns to remove the attribute while maintaining the identity of b_y.

Our main focus is on the attribute transfer module. F is an auxiliary part to maintain identity information of a_y during the transfer. Ideally, the output of G as a_y can be used as an example to be transferred to the output of F as b, and transferring attributes from a_y to b should generate b_y exactly.

Attribute Transfer Module. The target image a and source image b_y often vary in pose, expression, skin colors and lighting. Taking b_y and a directly into CNNs cannot produce faithful results because some attributes are not aligned well. To address the differences in pose and expression, we introduce a flow sub-network; and for the difference in lighting conditions and skin colors, we devise a refinement sub-network. The networks are shown in Figure 3. These two networks are trained end-to-end.

Figure 4: Step-by-step results of geometry-aware flow. The columns are respectively the target image a, source image b_y, warped source image b'_y, result a_y without flow warping and full model's result a_y. Without the geometry-aware flow, attributes between the source image and target image are not aligned correctly.
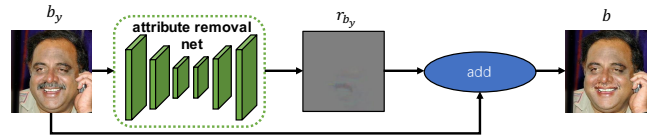

Figure 5: Network architecture of attribute removal module.

- **Flow Sub-Network**. The flow sub-network learns to generate the geometry-aware flow and an attention mask. The geometry-aware flow is generated as $\{\Phi^x, \Phi^y\}$, and the source image b_y is warped as b'_y. The output pixel of the warped image at location (i, j) is given as

$$b'_y(i, j) = \sum_{(i',j') \in N} (1 - |i + \Phi^x(i, j) - i'|) \\ (1 - |j + \Phi^y(i, j) - j'|)b_y, \tag{1}$$

where N stands for 4-pixel neighbors of $(x + \Phi^x, y + \Phi^y)$. We note that b'_y is differentiable to $\{\Phi^x, \Phi^y\}$ (Liu et al. 2017). Thus the flow sub-network can be trained end-to-end and integrated seamlessly into the proposed pipeline. An attention mask m is learned to select pixels from a and b'_y, and the result is blended as $a^*_y = m * a + (1 - m) * b'_y$. The geometry-aware flow is learned to align poses and expressions between a and b_y. As shown in Figure 4, without flow warping the mustache and eyeglasses are misaligned and the method produces erroneous results. More results are shown in Figure 7 in the experiments section.

- **Refinement Sub-Network**. Given a^*_y as input, the refinement sub-network learns to synthesize an appearance residual r_{a_y} to compensate for differences in skin colors and lightings. The final result is generated as $a_y = \alpha r_{a_y} + a^*_y$, where α is a hyper parameter to control the balance between flow sub-network and refinement sub-network. As shown in Figure 8, when skin colors or lighting conditions are different, the sub-network can generate appearance residuals to fix these differences.

Attribute Removal Module. For the attribute removal module, the input b_y and output b are of the same pose and expression, so we do not need the flow sub-network to warp input image. We adapt an encoder-decoder structure shown in Figure 5. To circumvent information loss, we use an U-Net architecture in which the $i - th$ layer is concatenated to the $(L-i) - th$ layer via skip connections. The network generates an image residual r_{b_y} and adds it onto b_y to get the final image, *i.e.*, $b = b_y + r_{b_y}$.

Multi-Objective Learning

Our full learning objective consists of four terms: adversarial, domain classification, reconstruction and geometry-aware flow terms. We detail these terms as follows.

Adversarial Objective. To make the images generated by attribute transfer network indistinguishable from real images, we adopt an adversarial loss. To stablize the training process and avoid common training failure modes such as mode collapse and gradient vanishing problem, we replace the original GAN loss as LSGAN (Mao et al. 2017) objective:

$$
\begin{aligned}
L_{adv-g} = {} & \mathbb{E}_{a \in A, b_y \in B} \big[\big\| 1 - D_{src}(G(a, b_y)) \big\|^2\big] \\
& + \mathbb{E}_{a \in A} \big[\big\| D_{src}(b_y) \big\|^2\big],
\end{aligned}
\tag{2}
$$

where the discriminator D attempts to discriminate between the real samples and the generated samples, and the attribute transfer network G aims to generate images that cannot be distinguished by the adversary.

We also apply an adversarial loss to encourage attribute removal network, F, to generate images indistinguishable from the faces without certain attribute.

$$
L_{adv-f} = \mathbb{E}_{b_y \in B} \big[\big\| 1 - D_{src}(F(b_y)) \big\|^2\big] + \mathbb{E}_{a \in A} \big[\big\| F(a) \big\|^2\big].
\tag{3}
$$

Domain Classification Objective. We also introduce an auxiliary classifier to allow our discriminators to constrain the result of attribute transfer and removal to falling within the correct attribute class domains. We formulate the constrain as a binary classification problem, since we only have two domains: with or without an attribute. The objective can be decomposed into two terms, namely a domain classification loss of real images, which is used to optimize D, and a domain classification loss of fake images that is used to optimize G. The former is defined as

$$
L_{cls}^r = \mathbb{E}_{a \in A}[\log P(c = 0|a)] + \mathbb{E}_{b_y \in B}[\log P(c = 1|b_y)].
\tag{4}
$$

The latter is

$$
\begin{aligned}
L_{cls}^f = {} & \mathbb{E}_{a \in F(b_y)}[\log P(c = 0|a)] \\
& + \mathbb{E}_{a_y \in G(a, b_y)}[\log P(c = 1|a_y)].
\end{aligned}
\tag{5}
$$

Reconstruction Objective. To ensure that the identity of a face with transferred attribute is preserved, we introduce and attribute removal network as a constraint. Intuitively, if we apply attribute transfer to a and remove it, then we should get back the original image. Likely, if we remove the attribute on b_y to a as a_y and then transfer back, we should get

the same image as b_y.

$$
\begin{aligned}
L_{rec} = {} & \mathbb{E}_{b_y \in B, a \in A}\big[\big\| F(G(a, b_y)) - a \big\|_1\big] \\
& + \mathbb{E}_{a_y \in G(a, b_y), b \in F(b_y)}\big[\big\| G(b, a_y) - b_y \big\|_1\big].
\end{aligned}
\tag{6}
$$

Geometry-Aware Flow Objective. Recall that the generated geometry-aware flow is used to warp a given reference image to align with a target image. Towards this goal, we introduce a landmark loss as well as TV regularization loss to facilitate the training of flow. For landmark loss, we use FAN (Bulat and Tzimiropoulos 2017) to detect 68 landmarks $\{x_j^{b_y}, y_j^{b_y}|_{j=1}^{68}\}$ for b_y and $\{x_j^a, y_j^a|_{j=1}^{68}\}$ for a. We require that the landmarks of b_y and a should be close, so the landmark loss is defined as

$$
\begin{aligned}
L_{lm} = {} & \sum\nolimits_{j=1}^{68} (\Phi^x(x_j^{b_y}, y_j^{b_y}) + x_j^{b_y} - x_j^a)^2 \\
& + (\Phi^y(x_j^{b_y}, y_j^{b_y}) + y_j^{b_y} - y_j^a)^2.
\end{aligned}
\tag{7}
$$

As the landmark loss can only be imposed on 68 landmarks, we hope that geometry-aware flow should be spatially smooth so that the structure of target image can be maintained. Thus, a TV regularizer is used. It is defined as

$$
L_{tv} = \|\nabla\Phi^x\|^2 + \|\nabla\Phi^y\|^2
\tag{8}
$$

We combine the landmark loss with TV regularizer and define the flow loss as

$$
L_{flow} = l_{lm} + l_{tv}.
\tag{9}
$$

Overall Objective. We combine the adversarial, domain classification, reconstruction, and flow objectives to obtain the overall objective as

$$
L_{full} = L_{adv-f} + L_{adv-g} + L_{cls}^r + L_{cls}^f + L_{rec} + L_{flow}.
\tag{10}
$$

This overall objective is optimized in a multi-task learning manner. Network hyper-parameters and loss weights are determined on a probe validation subset.

Experiments

In this section, we comprehensively evaluate our approach on different benchmarks with dedicated metrics.

Implementation Details

Training Details. Input image values are normalized to $[-1, 1]$. All models are trained using Adam (Kingma and Ba 2014) optimizer with a base learning rate of 0.002, and a batch size of 8. We perform data augmentation by random horizontal flipping with a probability of 0.5.

Network Architecture. Our network architecture is inspired by Pix2PixHD (Wang et al. 2018). There are four encoder-decoder networks in our pipeline including the flow sub-network, attention mask sub-network, refinement sub-network and attribute removal sub-network. They share similar architectures as U-Net and are composed of three convolutional layers, three residual layers and three deconvolutional layers. For the discriminator we adopt Patch-GAN (Isola et al. 2017), which perform real and fake clasification on local image patches.
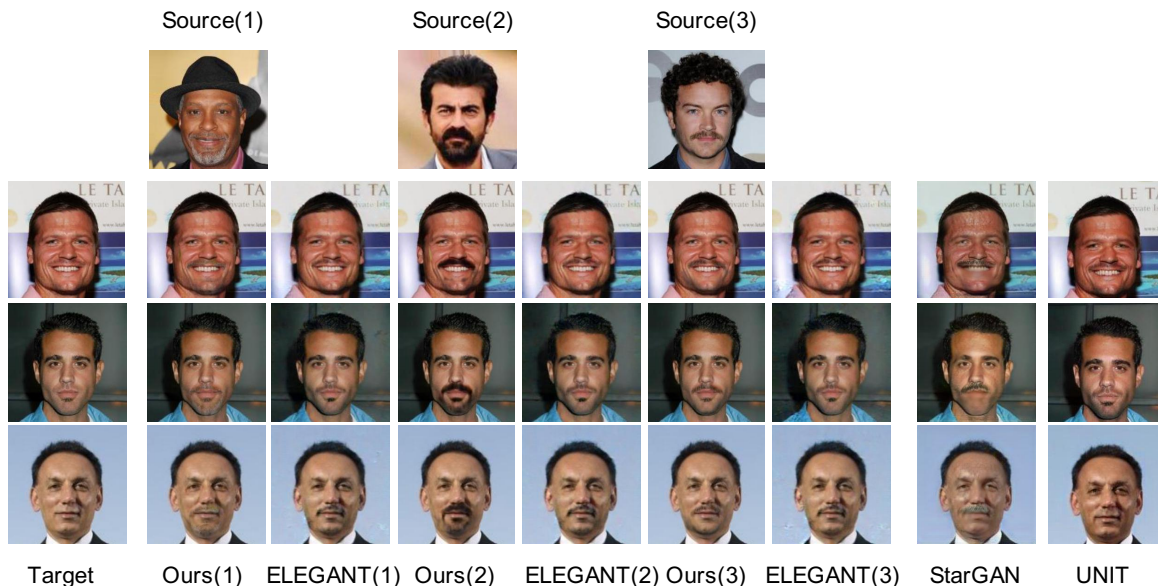
Figure 6: Mustache transfer results on the CelebA dataset. The first six columns the results of the proposed method and ELE-GANT (Xiao, Hong, and Ma 2018) for three different source images. The last two columns show the result of StarGAN (Choi et al. 2017) and UNIT (Liu, Breuel, and Kautz 2017), respectively.

Competing Methods. We choose state-of-the-art Star-GAN (Choi et al. 2017), UNIT (Liu, Breuel, and Kautz 2017) and ELEGANT (Xiao, Hong, and Ma 2018) as our baselines. StarGAN and UNIT both perform semantic-level facial attribute manipulation, while ELEGANT is capable of transferring instance-level attributes based on given source image.

Datasets

CelebA (Liu et al. 2015) is a large-scale face attributes dataset containing more than 200000 images of celebrity annotated with 40 attributes. We use the standard training, validation and test splits. For preprocessing, we crop images to 178x178, and resize them to 256x256.

CelebA-HQ (Karras et al. 2017) is a higher quality version of CelebA dataset that allows experimentation with up to 1024x1024 resolution. We only use this dataset as an input to genetate high-resolution results and do not use it for training.

Comparisons to Prior Works

The comparison is performed on three aspects, including attribute-level face manipulation, instance-level attribute transfer and distribution-level evaluation. We name our approach as GeoGAN for easy reference.

Attribute-level Face Manipulation. To evaluate the ability of a method in manipulating a desired attribute, we examined the classification accuracy of synthesized images. We trained three binary facial attribute classifiers for attributes including *Eyeglass, Mustache, Goatee* separately on the CelebA dataset. Using a ResNet-18(He et al. 2016) architecture, we achieved above 95% accuracy for each of these classifiers. We then trained each of image transla-

tion models using the same training set and performed face attribute manipulation on unseen validation set. For ELE-GANT and our method, we sampled exemplar images from the same validation set with inverse label. Finally we classified the attribute of these manipulated images using the above-mentioned classifiers. The result is shown in Table 1. Though StarGAN achieved the best classification accuracy, the quality of their generated images is poor and all attributes follow the same pattern, as shown in Figure 6. Our model beats ELEGANT and UNIT at a large margin, suggesting that our model can manipulate attributes more accurately.

Instance-level Evaluation. As StarGAN and UNIT cannot perform attribute transfer at instance level, we compared our method with ELEGANT on this task. To evaluate the faithfulness, we introduced faithfulness score.

The faithfulness score is designed as follows. Given a target image a and source image b_y, the region of desired attribute is cropped as a^{attr} and b_y^{attr} according to facial landmarks. We then extract the features of these cropped regions as f_a and f_{b_y} using VGG18 (Simonyan and Zisserman 2014) pretrained on ImageNet(Deng et al. 2009). Finally the faithfulness score s_{faith} is computed as the distance of two cropped regions on normalized feature space:

$$s_{faith} = \left\| \frac{f_a}{\|f_a\|_2} - \frac{f_{b_y}}{\|f_{b_y}\|_2} \right\|_2 \qquad (11)$$

Lower score indicates more faithful transfer. As shown in Table 1, our method outperforms ELEGANT in term of faithfulness at a large margin. This is because our model introduced flow sub-network to warp pixels directly from target image thus increasing the faithfulness and sharpness of the synthesized image.

Metric		FID Score			Faithfulness Score			Attribute Cls. Accuracy(%)		
Attribute		Eyeglasses	Mustache	Goatee	Eyeglasses	Mustache	Goatee	Eyeglasses	Mustache	Goatee
Semantic-level Transfer	UNIT	*0.316*	*0.356*	*0.321*	\	\	\	89.5	51.9	61.6
	StarGAN	0.386	0.390	0.374	\	\	\	*99.6*	*98.8*	*98.8*
Instance-level Transfer	ELEGANT	0.410	0.387	0.378	0.946	0.915	0.905	86.8	82.1	82.3
	GeoGAN (-F)	0.355	0.366	0.352	0.959	0.883	0.855	**96.0**	**90.2**	81.5
	GeoGAN (-H)	0.353	0.337	0.324	**0.786**	**0.825**	**0.810**	80.7	78.9	77.4
	GeoGAN (full)	**0.351**	**0.336**	**0.324**	0.806	0.832	0.811	91.2	87.8	**91.9**

Table 1: Benchmarking results of different methods on the CelebA dataset, w.r.t. three metrics including the FID Score (lower is better), Faithfulness Score (lower is better) and Attribute Classification Accuracy (higher is better), on both semantic-level and instance-level tracks. The italic entries indicate the best performance in the Semantic-level Transfer track while the bold entries indicated the best performance in the Instance-level Transfer track.

Method	Attribute Quality	Perceptual Realism
StarGAN	11.3%	11.9%
UNIT	26.1%	29.3%
ELEGANT	17.4%	16.7%
GeoGAN (ours)	**45.2%**	**42.1%**

Table 2: User study of different methods (the higher the better) w.r.t. both attribute quality and perceptual realism. Each column represents user preferences that sum to 100%.

Distribution-level Evaluation. To measure the quality of generated images from different models quantitatively, we calculated Fréchet Inception Distance (FID)(Heusel et al. 2017) between real images and generated images to measure the quality of generated images. The lower the score, the better the result. As shown in Table 1, our full model performs better than StarGAN and ELEGANT in all attributes. The FID score of UNIT is relatively low because many generated images look alike the original image and the desired attribute is not manipulated accurately. This is indicated by the low classification accuracy in Table 1.

Figure 6 shows examples of images generated when transferring different attributes: the generated images of our model are of high visual quality. These generated images confirm that our model not only preserves the identity of original image, but also captures the attribute to be transferred from source image. Though StarGAN and UNIT have the capability of performing attribute manipulation, many of their results are blurry. Importantly, they can only generate a homogeneous style of attribute for many input images, which lack diversity.

User Study. We performed a user survey to assess our model in terms of attribute manipulation and visual quality. Given an input image, the user were required to choose the best generated image based on two criteria: quality of transfer in attributes and perceptual realism. The options were four randomly shuffled images generated from different methods of

Figure 7: Ablation results of the flow sub-network. From left to right each column represents the target image, source image, our result without flow sub-network and our full result.

the same identity. The modified attribute is selected among *Eyeglasses, Mustache, Goatee* equally. As is shown in Table 2, our model obtained the majority of votes for the best visual quality while preserving the desired attribute.

Ablation Study

In the ablation study, we considered three variants of our model: (i) GeoGAN (full): our full model. (ii) GeoGAN (-F): our model with the flow sub-network of attribute transfer module removed, and (iii) GeoGAN (-H): our model with the refinement sub-network of attribute transfer module removed. Table 1 lists the FID score, faithfulness score and attribute classification accuracy of these variants. The results suggest the two networks contribute complementary to our full model. More discussion are given below.

Effectiveness of Geometry-aware Flow. Without the flow sub-network, the performance in FID score and faithfulness score dropped significantly, suggesting that the flow sub-network is essential towards faithful facial attribute transfer. Figure 7 provides examples of our model without flow sub-
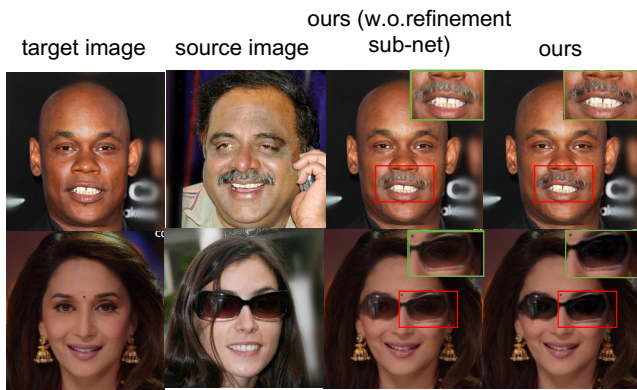
Figure 8: Ablation results of the refinement sub-network. From left to right each column represents target image, source image, our result without refinement sub-network and our full result.
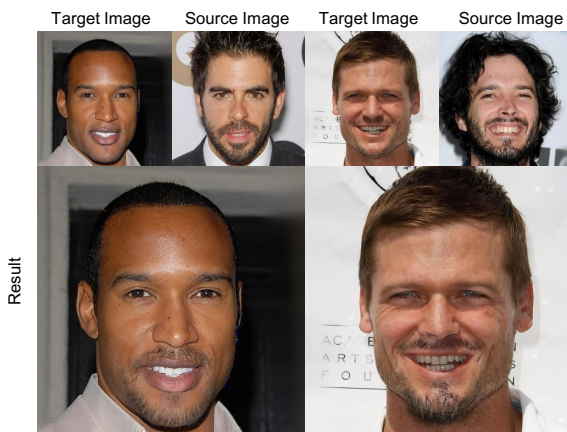

Figure 9: Each triplet includes a source image, target image and our high-res result.

network. Without the flow sub-network, synthesized images are blurry and do not preserve features in the source image. **Effectiveness of Appearance Residual.** Without the refinement sub-network, a drastic drop of the performance in classification is observed. As shown in Figure 8, without the refinement sub-network the flow sub-network cannot cope with differences in skin colors and lighting conditions.

Results on High-Resolution Images

As the geometry-aware flow is invariant with regard to different resolutions, our model is able to produce high-resolution results without re-training. The result is shown in Figure 9. Given a low-resolution source image and target image, the result is generated by upsampling generated geometry-aware flow, attention mask and appearance residual then apply them on the source image and target image at original resolution. Benefited from the flow sub-network, our model preserves high-frequency patterns in source image and target image well. Scaling other baselines to producing high-resolution results is much harder in comparison
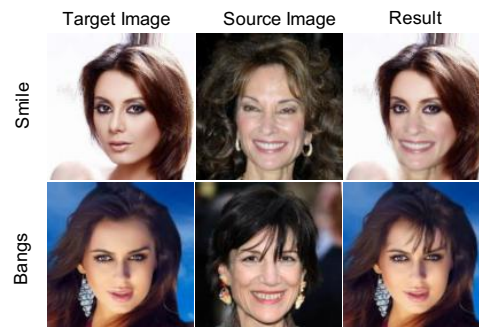

Figure 10: Experiment results on non-local attributes such as smiles and bangs. Every row contains results for one attribute.

to our approach as training on high-resolution images are needed, and the process will be extremely time-consuming and computational ineffective.

Results on More Attributes

As shown in Figure 10, our method generalizes well to other challenging attributes such as bangs and smiling which exhibit a high degree of non-local displacement. For "smiling", the flow subnetwork first warped and blended the teeth and cheek regions into target image, then the refinement subnetwork compensated for the local mis-alignments along the lip. On another non-local attribute "bangs", our flow network well transferred the bang from the source face to the target face. The global traits such as hair color were addressed by our refinement network.

Conclusion

We have proposed the notion of geometry-aware flow to address the problem of instance-level facial attribute transfer. In comparison to prior arts, our method has several appealing properties: (1) Geometry-aware flow serves as a well-suited representation for modeling the transformation between instance-level facial attributes, despite of the large pose gap between the source and target face images. (2) When combined with appearance residual produced by the refinement sub-network, our approach is capable of handling potential appearance gap between the source and the target. (3) Geometry-aware flow can be readily applied to high-resolution face images and generate desired attributes with realistic details.

Though our current framework handles one attribute at a time, it can be readily extended to handle multiple attributes in one pass. For example, following the idea in 'StyleBank' (Chen et al. 2017), the flow, mask subnetworks and attribute removal network in our framework can be augmented to output multiple flows, masks and image residuals as an attribute bank. During multi-attribute inference, only one element in the 'attribute banks' is activated. It is definitely an interesting direction to explore as our future work.

References

Azadi, S.; Pathak, D.; Ebrahimi, S.; and Darrell, T. 2018. Compositional gan: Learning conditional image composition. *CoRR* abs/1807.07560.

Brock, A.; Lim, T.; Ritchie, J. M.; and Weston, N. 2016. Neural photo editing with introspective adversarial networks. *CoRR* abs/1609.07093.

Bulat, A., and Tzimiropoulos, G. 2017. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*.

Chen, D.; Yuan, L.; Liao, J.; Yu, N.; and Hua, G. 2017. Stylebank: An explicit representation for neural image style transfer. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2770–2779.

Choi, Y.; Choi, M.-J.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2017. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *CoRR* abs/1711.09020.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. *CVPR*.

Gardner, J. R.; Kusner, M. J.; Li, Y.; Upchurch, P.; Weinberger, K. Q.; and Hopcroft, J. E. 2015. Deep manifold traversal: Changing labels with convolutional features. *CoRR* abs/1511.06421.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. *CVPR*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*.

Huang, C.; Li, Y.; Loy, C. C.; and Tang, X. 2018. Deep imbalanced learning for face recognition and attribute prediction. *arXiv preprint arXiv:1806.00194*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. *CVPR*.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; and Kavukcuoglu, K. 2015. Spatial transformer networks. In *NIPS*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *CoRR* abs/1710.10196.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *CoRR* abs/1312.6114.

Lample, G.; Zeghidour, N.; Usunier, N.; Bordes, A.; Denoyer, L.; and Ranzato, M. 2017. Fader networks: Manipulating images by sliding attributes. In *NIPS*.

Lin, C.-H.; Yumer, E.; Wang, O.; Shechtman, E.; and Lucey, S. 2018. St-gan: Spatial transformer generative adversarial networks for image compositing. *CoRR* abs/1803.01837.

Liu, M.-Y., and Tuzel, O. 2016. Coupled generative adversarial networks. In *NIPS*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*.

Liu, Z.; Yeh, R. A.; Tang, X.; Liu, Y.; and Agarwala, A. 2017. Video frame synthesis using deep voxel flow. In *ICCV*.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *NIPS*.

Loy, C. C.; Luo, P.; and Huang, C. 2017. Deep learning face attributes for detection and alignment. In *Visual Attributes*. Springer. 181–214.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y. K.; Wang, Z.; and Smolley, S. P. 2017. Least squares generative adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)* 2813–2821.

Nguyen, M. H.; Lalonde, J.-F.; Efros, A. A.; and la Torre, F. D. 2008. Image-based shaving. *Computer Graphics Forum*.

Perarnau, G.; van de Weijer, J.; Raducanu, B.; and Álvarez, J. M. 2016. Invertible conditional gans for image editing. *CoRR* abs/1611.06355.

Shen, W., and Liu, R. 2017. Learning residual images for face attribute manipulation. *CVPR*.

Shu, Z.; Yumer, E.; Hadap, S.; Sunkavalli, K.; Shechtman, E.; and Samaras, D. 2017. Neural face editing with intrinsic image disentangling. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 5444–5453.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Xiao, T.; Hong, J.; and Ma, J. 2018. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. *CoRR* abs/1803.10562.

Yeh, R. A.; Liu, Z.; Goldman, D. B.; and Agarwala, A. 2016. Semantic facial expression editing using autoencoded flow. *CoRR* abs/1611.09961.

Zhou, T.; Tulsiani, S.; Sun, W.; Malik, J.; and Efros, A. A. 2016. View synthesis by appearance flow. In *ECCV*.

Zhou, S.; Xiao, T.; Yang, Y.; Feng, D.; He, Q.; and He, W. 2017. Genegan: Learning object transfiguration and attribute subspace from unpaired data. *CoRR* abs/1705.04932.