Supplementary Material: Toward a Controllable Disentanglement Network

Zengjie Song[1], Oluwasanmi Koyejo[2], Jiangshe Zhang[1]
[1]School of Mathematics and Statistics, XJTU, Xi'an 710049, China
[2]Department of Computer Science, UIUC, Urbana, IL 61801 USA
zjsong@hotmail.com, sanmi@illinois.edu, jszhang@mail.xjtu.edu.cn

1 Network Architectures

The network architecture of the CDNet consists of four parts: two encoders $\text{Enc}_{\mathbf{y}}$ and $\text{Enc}_{\mathbf{z}}$, decoder Dec/generator Gen, and discriminator Dis. The $\text{Enc}_{\mathbf{y}}$ and the $\text{Enc}_{\mathbf{z}}$ have the similar architecture, except the dimensionality and the activation function of the last fully-connected layer. The architecture details of these four modules are shown in Tables 1 and 2, where symbols "Conv", "FC", "Concat", and "Deconv" denote convolution, fully-connected, concatenation, and deconvolution operations, respectively.

Table 1: Details of the network architecture used for MNIST dataset.

Module	Operation	Kernel	Stride	Padding	Filters	BN	Activation	Dropout
$\text{Enc}_{\mathbf{y}}$ and $\text{Enc}_{\mathbf{z}}$	Conv	4×4	2×2	1×1×1×1	32	✔	Leaky ReLU ($a = 0.2$)	✘
	Conv	4×4	2×2	1×1×1×1	64	✔	Leaky ReLU ($a = 0.2$)	✘
	FC	-	-	-	1000	✘	Leaky ReLU ($a = 0.2$)	✔ ($p = 0.5$)
	FC	-	-	-	$\text{Enc}_{\mathbf{y}}$: 10 $\text{Enc}_{\mathbf{z}}$: 10	✘	$\text{Enc}_{\mathbf{y}}$: Softmax $\text{Enc}_{\mathbf{z}}$: Linear	✘
Dec or Gen	Concat	Concatenate $\hat{\mathbf{y}}$ and $\mathbf{z}$ on 1st dimension						
	FC	-	-	-	1000	✔	ReLU	✘
	Concat	Concatenate $\hat{\mathbf{y}}$ and last layer's output on 1st dimension						
	FC	-	-	-	3136	✔	ReLU	✘
	Concat	Replicate $\hat{\mathbf{y}}$ and append as additional constant input channels						
	Deconv	4×4	2×2	1×1×1×1	32	✔	ReLU	✘
	Concat	Replicate $\hat{\mathbf{y}}$ and append as additional constant input channels						
	Deconv	4×4	2×2	1×1×1×1	1	✘	Sigmoid	✘
Dis	Conv	5×5	1×1	2×2×2×2	32	✘	Leaky ReLU ($a = 0.2$)	✘
	Conv	4×4	2×2	1×1×1×1	64	✔	Leaky ReLU ($a = 0.2$)	✘
	Conv	4×4	2×2	1×1×1×1	128	✔	Leaky ReLU ($a = 0.2$)	✘
	FC	-	-	-	128	✘	Leaky ReLU ($a = 0.2$)	✔ ($p = 0.5$)
	FC	-	-	-	1	✘	Sigmoid	✘

Table 2: Details of the network architecture used for CelebA dataset.

Module	Operation	Kernel	Stride	Padding	Filters	BN	Activation	Dropout
$\text{Enc}_{\mathbf{y}}$ and $\text{Enc}_{\mathbf{z}}$	Conv	4×4	2×2	1×1×1×1	64	✔	Leaky ReLU ($a = 0.2$)	✘
	Conv	4×4	2×2	1×1×1×1	128	✔	Leaky ReLU ($a = 0.2$)	✘
	Conv	4×4	2×2	1×1×1×1	256	✔	Leaky ReLU ($a = 0.2$)	✘
	FC	-	-	-	4000	✘	Leaky ReLU ($a = 0.2$)	✔ ($p = 0.5$)
	FC	-	-	-	2000	✘	Leaky ReLU ($a = 0.2$)	✔ ($p = 0.5$)
	FC	-	-	-	$\text{Enc}_{\mathbf{y}}$: 40 $\text{Enc}_{\mathbf{z}}$: 1000	✘	$\text{Enc}_{\mathbf{y}}$: Sigmoid $\text{Enc}_{\mathbf{z}}$: Linear	✘
Dec or Gen	Concat	Concatenate $\hat{\mathbf{y}}$ and $\mathbf{z}$ on 1st dimension						
	FC	-	-	-	2000	✔	ReLU	✘
	Concat	Concatenate $\hat{\mathbf{y}}$ and last layer's output on 1st dimension						
	FC	-	-	-	4000	✔	ReLU	✘
	Concat	Concatenate $\hat{\mathbf{y}}$ and last layer's output on 1st dimension						
	FC	-	-	-	16384	✔	ReLU	✘
	Concat	Replicate $\hat{\mathbf{y}}$ and append as additional constant input channels						
	Deconv	4×4	2×2	1×1×1×1	128	✔	ReLU	✘
	Concat	Replicate $\hat{\mathbf{y}}$ and append as additional constant input channels						
	Deconv	4×4	2×2	1×1×1×1	64	✔	ReLU	✘
	Concat	Replicate $\hat{\mathbf{y}}$ and append as additional constant input channels						
	Deconv	4×4	2×2	1×1×1×1	3	✘	Tanh	✘
Dis	Conv	5×5	1×1	2×2×2×2	32	✘	Leaky ReLU ($a = 0.2$)	✘
	Conv	4×4	2×2	1×1×1×1	128	✔	Leaky ReLU ($a = 0.2$)	✘
	Conv	4×4	2×2	1×1×1×1	256	✔	Leaky ReLU ($a = 0.2$)	✘
	Conv	4×4	2×2	1×1×1×1	256	✔	Leaky ReLU ($a = 0.2$)	✘
	FC	-	-	-	512	✘	Leaky ReLU ($a = 0.2$)	✔ ($p = 0.5$)
	FC	-	-	-	1	✘	Sigmoid	✘

2 Additional Results

We provide two groups of experiments to further illustrate the effectiveness of our CDNet model. The first group of experiments are synthesizing face images with several target facial attributes successively (see Section 2.1). The second group of experiments are synthesizing face images with the specific facial attribute and, simultaneously, with the designated attribute intensities (see Section 2.2). All experiments are conducted on the CelebA test set. And three relevant models, i.e., AE-XCov, IcGAN, and VAE/GAN, are employed as the baselines.

2.1 Disentanglement

The results are shown in Figures 1-5.

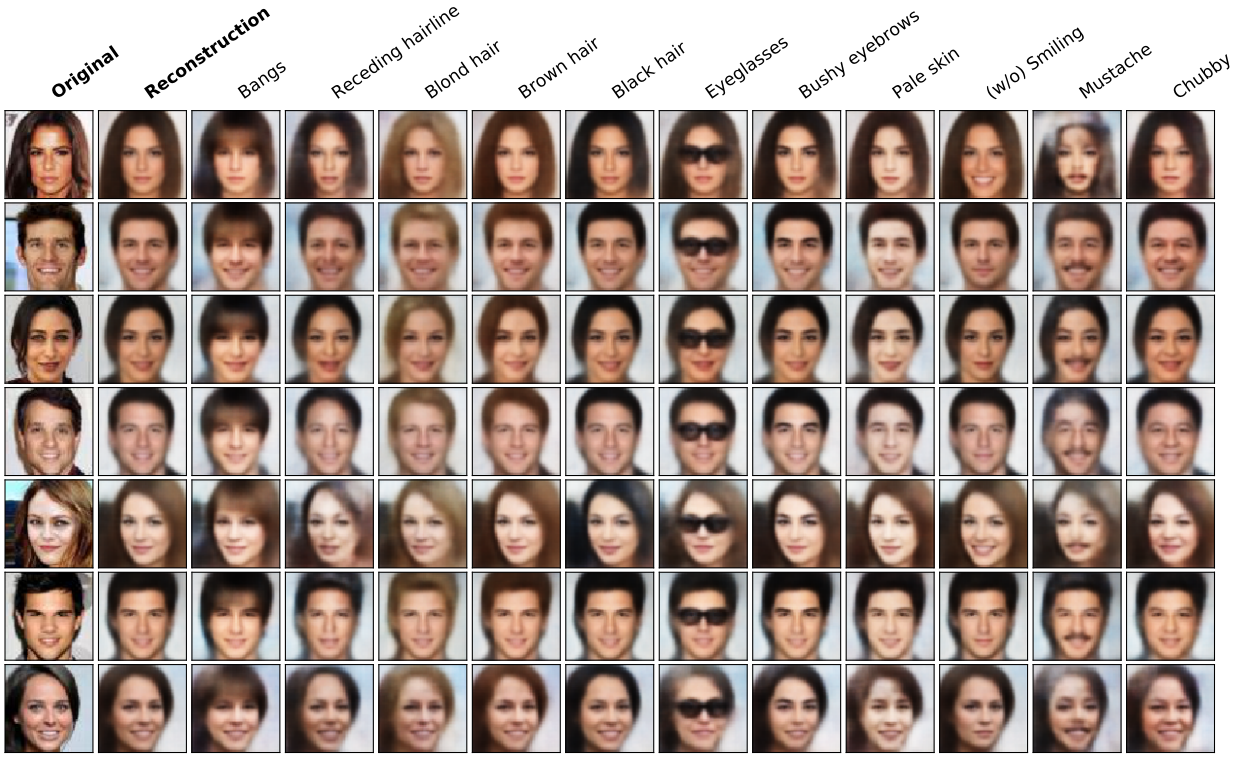


Figure 1: Synthesized face images with the designated attributes by the **AE-XCov** model. Best viewed in color.

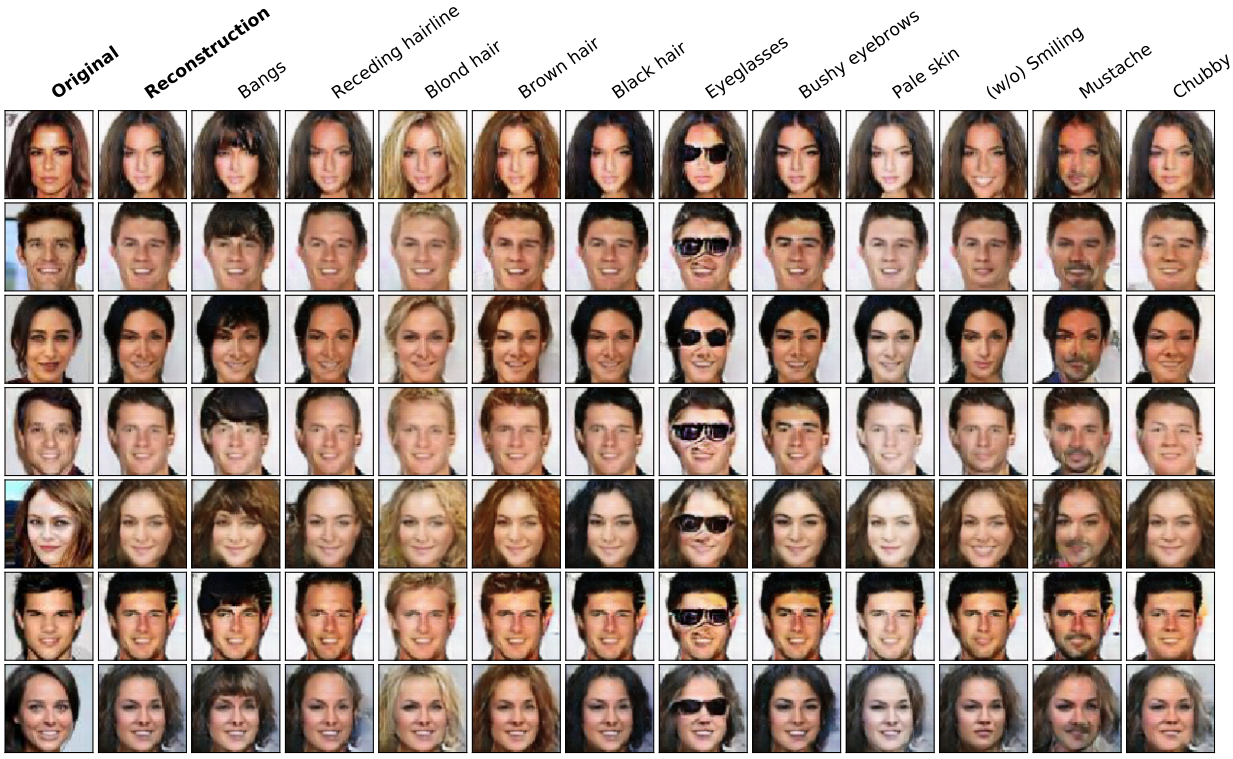


Figure 2: Synthesized face images with the designated attributes by the **IcGAN** model. Best viewed in color.

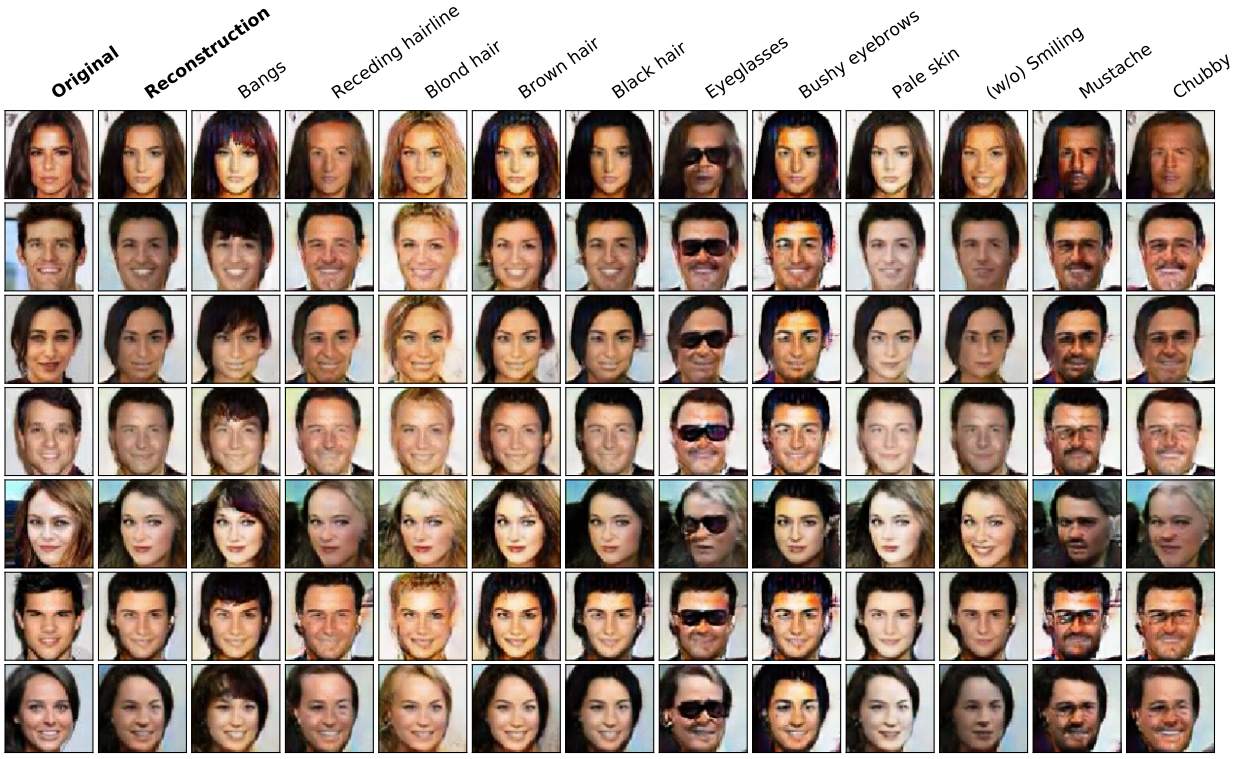


Figure 3: Synthesized face images with the designated attributes by the **VAE/GAN** model. Best viewed in color.

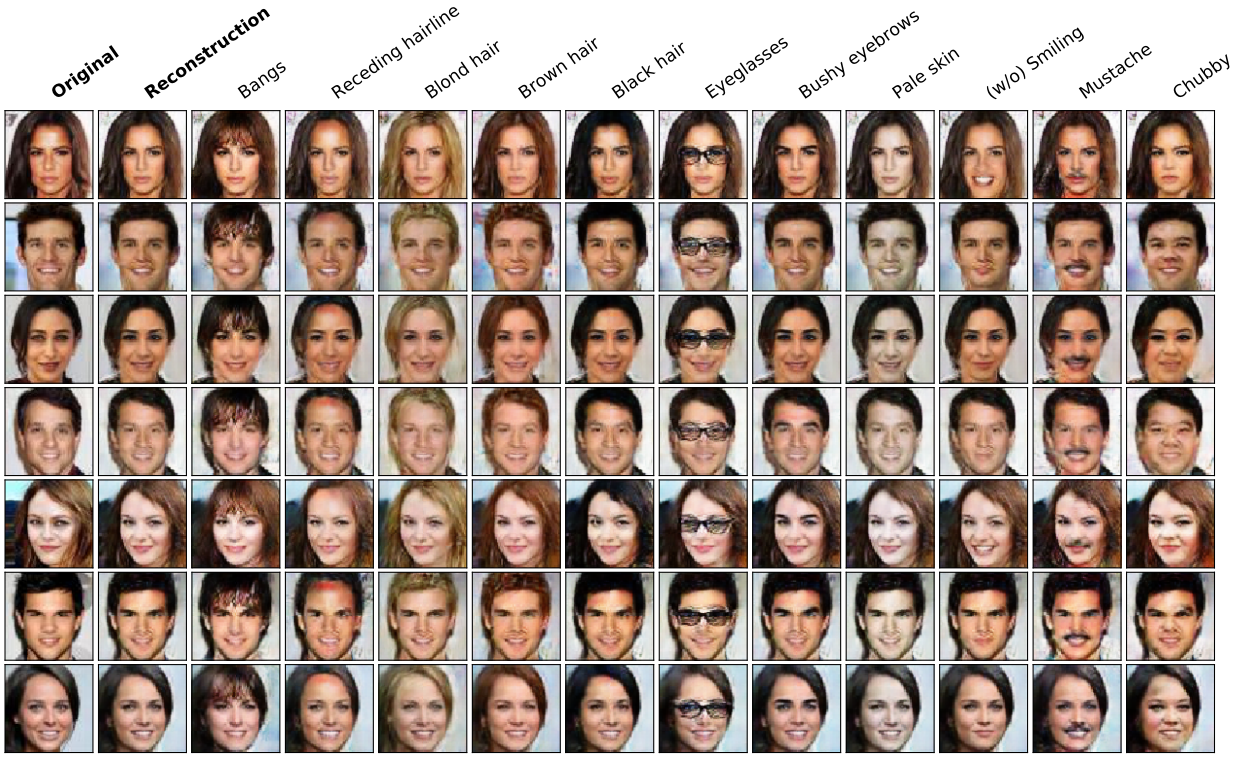


Figure 4: Synthesized face images with the designated attributes by the **CDNet-XCov** model. Best viewed in color.

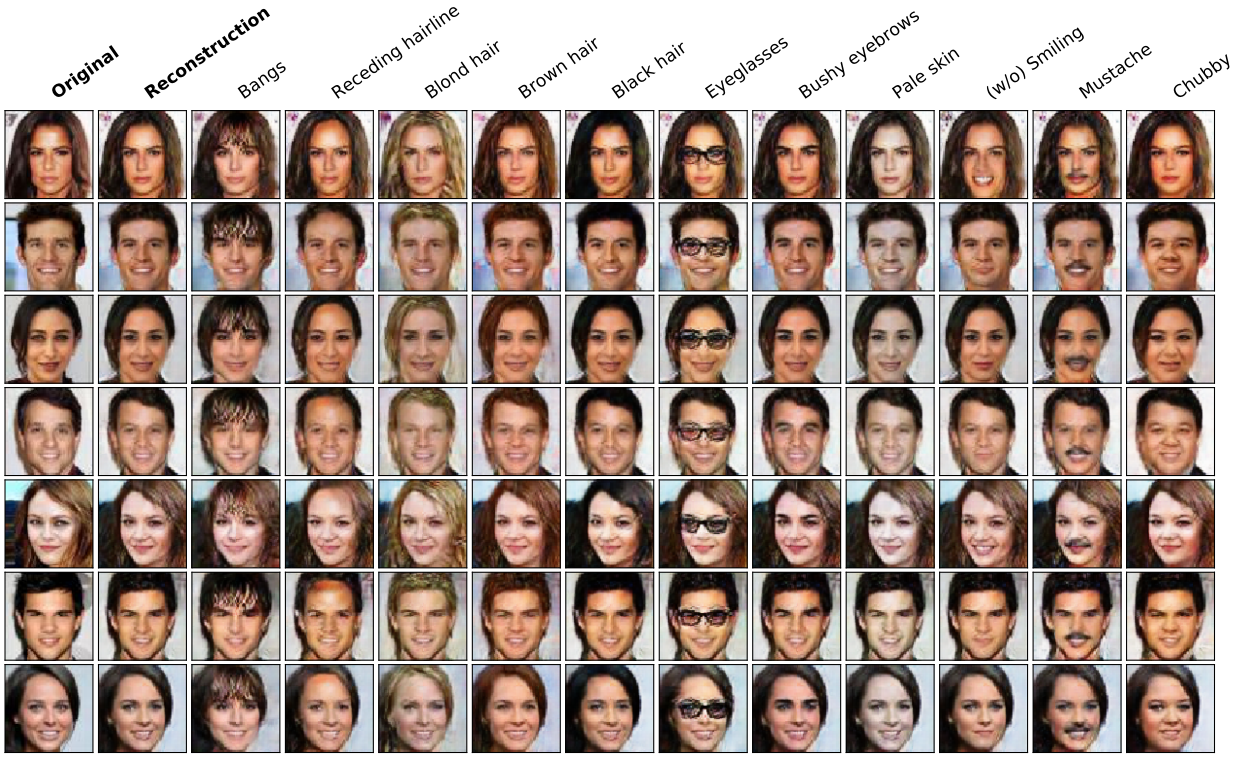
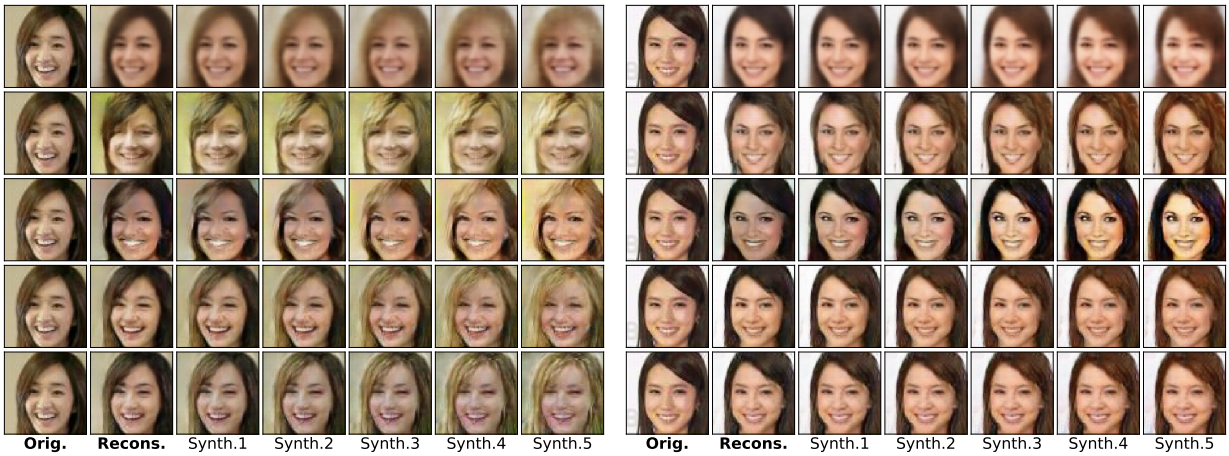


Figure 5: Synthesized face images with the designated attributes by the **CDNet-dCov** model. Best viewed in color.

2.2 Controllable Disentanglement

The results are shown in Figures 6 and 7.



(a) Blond hair

(b) Brown hair

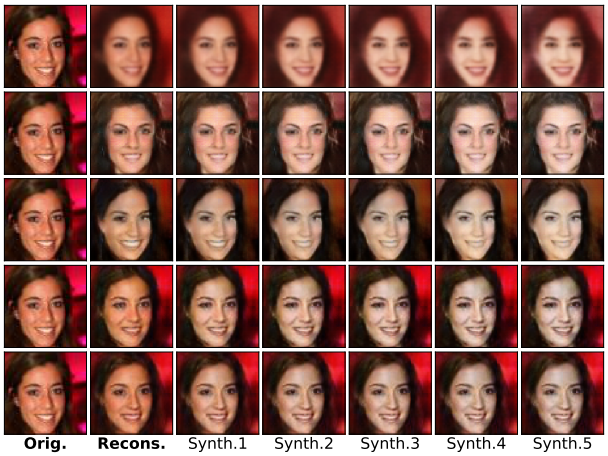
(c) Black hair

(d) Bangs

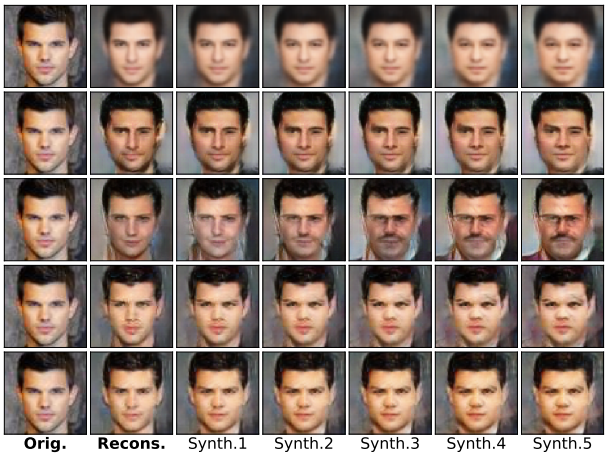
(e) Receding hairline

(f) Mustache

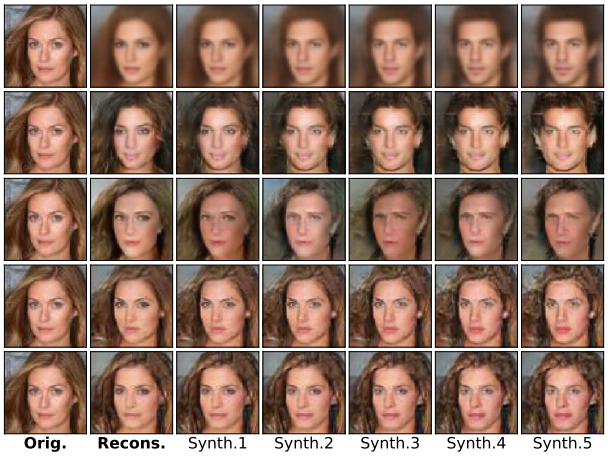
Figure 6: Synthesized face images with different facial attributes and attribute intensities (Part-1). The results in each panel, from the first row to the last row, are obtained by AE-XCov, IcGAN, VAE/GAN, CDNet-XCov, and CDNet-dCov, respectively. In each panel, the first column shows the original test image, the second column for reconstructions, and the remaining five columns for synthesized images with different attribute intensities, from weaker levels to stronger ones. Best viewed in color.



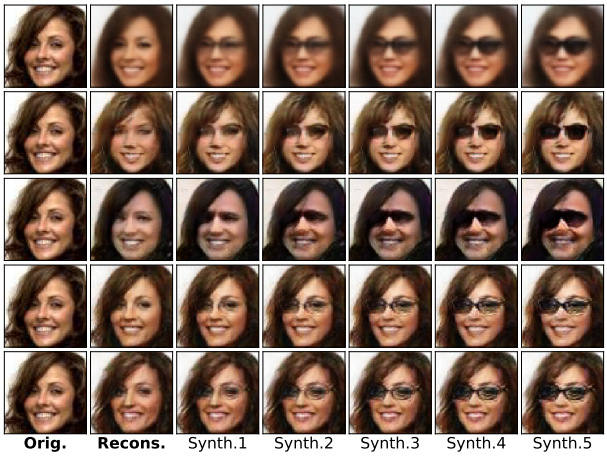
(a) Pale skin



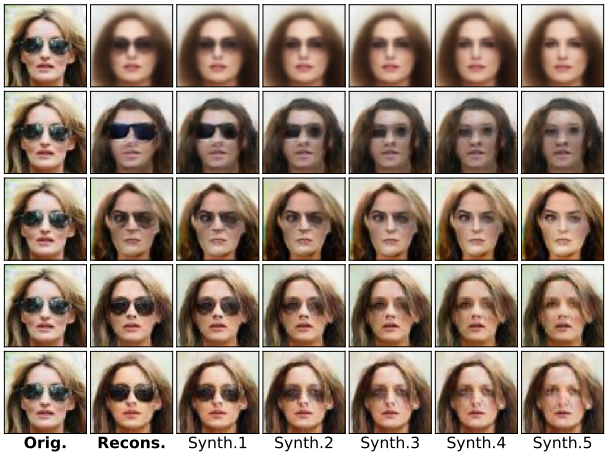
(b) Chubby



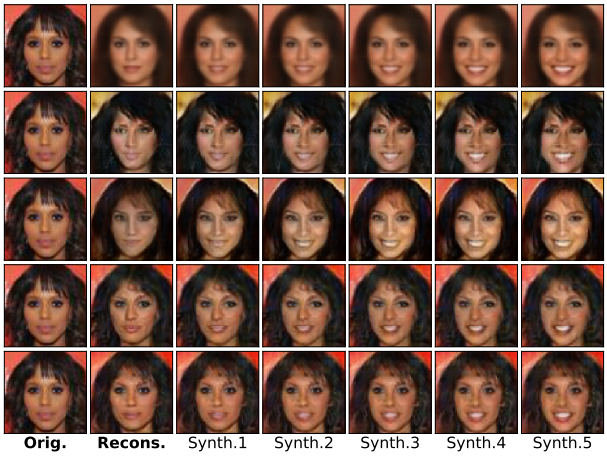
(c) Male



(d) Eyeglasses



(e) w/o Eyeglasses



(f) Smiling

Figure 7: Synthesized face images with different facial attributes and attribute intensities (Part-2). The results in each panel, from the first row to the last row, are obtained by AE-XCov, IcGAN, VAE/GAN, CDNet-XCov, and CDNet-dCov, respectively. In each panel, the first column shows the original test image, the second column for reconstructions, and the remaining five columns for synthesized images with different attribute intensities, from weaker levels to stronger ones. Best viewed in color.

3 Comparison with StarGAN

In this section, we qualitatively and quantitatively compare our model with one of the state-of-the-art multi-domain image translation models, i.e., StarGAN, on CelebA face image reconstruction and disentanglement tasks, respectively. Note that for face image editing on CelebA dataset, the original StarGAN was trained only with 5 facial attributes. In order to explore StarGAN's disentanglement performance on other attributes, we used the publicly released code[1] to retrain the model with all 40 facial attributes. The size of test images is 64×64, and for StarGAN all input images were upsampled to 128×128 while the output images were downsampled to 64×64 for comparison.

3.1 Reconstruction

The reconstructed face images are illustrated in Figure 8. As we can see from Figure 8, StarGAN performs better to depict local features such as hair and background textures. However, it cannot preserve the core object identity and the complexion very well compared with our CDNet models. In terms of the quality assessment of reconstructed images, as shown in Table 3, the two CDNet models also outperform the StarGAN across all three evaluation criteria. We conclude that the image reconstruction ability (similarly the disentanglement performance demonstrated later) of StarGAN heavily depends on the number of transferred domains (i.e., facial attributes in these experiments); and that StarGAN is more suitable for processing images with a small number of transferred domains (e.g., the original StarGAN has shown effectiveness of image editing with 5 facial attributes on CelebA dataset).

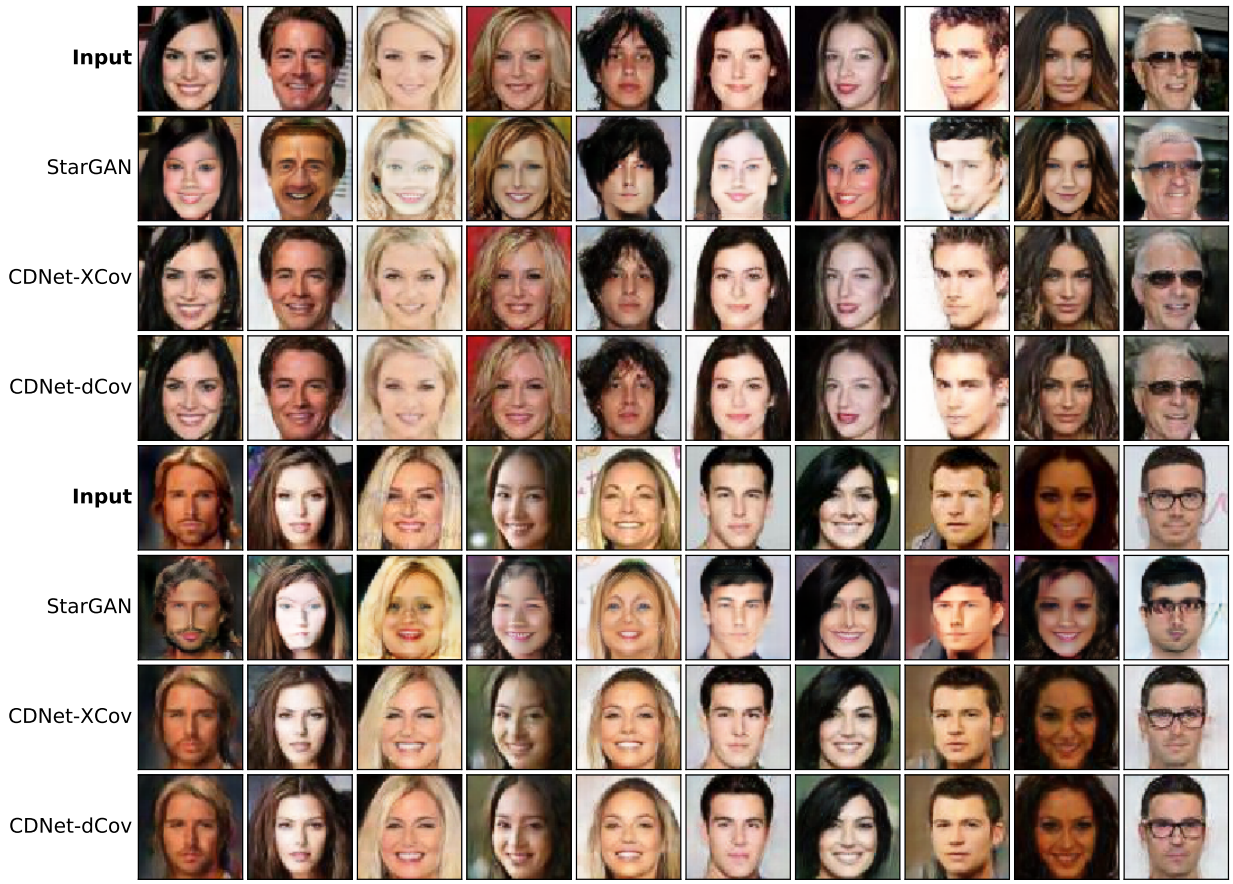


Figure 8: Reconstructions of face images from CelebA test set. Best viewed in color.

[1] Public PyTorch code of StarGAN: `https://github.com/yunjey/stargan`.

Table 3: Reconstruction quality on CelebA test set. The results are formatted as mean $\pm$ standard deviation.

Model	RMSE	PSNR	SSIM
StarGAN	0.1337 ± 0.0297	17.6781 ± 1.8341	0.7898 ± 0.0852
CDNet-XCov	0.0834 ± 0.0175	21.7564 ± 1.7683	0.9099 ± 0.0377
CDNet-dCov	0.0828 ± 0.0172	21.8181 ± 1.7580	0.9108 ± 0.0377

3.2 Disentanglement

To qualitatively compare the disentanglement ability of CDNet with that of StarGAN, we visualize some of the synthetic face images. As can be seen from Figures 9-11, although StarGAN is able to add the designated attributes into synthetic faces, the object identities of new faces are visibly inconsistent with the input ones. Besides, StarGAN shows limited ability to control degree of disentanglement, because there is no conspicuous visual difference among generated images with the same facial attribute (see Figures 10 and 11). For example, the attribute intensity of "Black hair" in StarGAN takes ten different values in set of $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9, 1.0, 1.5, 2.0\}$ in sequence, but the hair colors reflected in all ten synthetic images are almost the same (Figure 10(b)). Similar phenomena can be observed in other attribute cases, such as "Bangs" (Figure 10(a)) and "Brown hair" (Figure 10(d)).

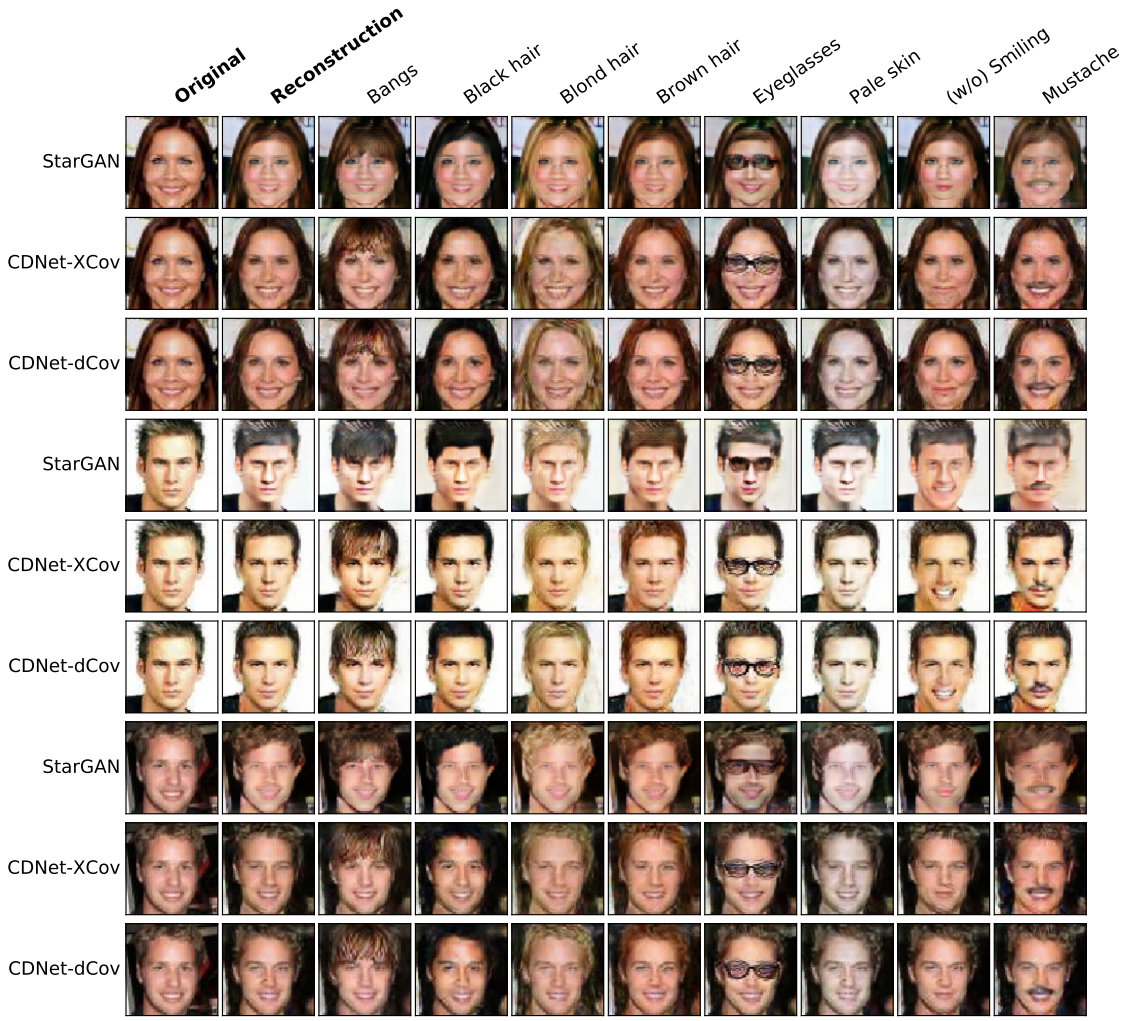
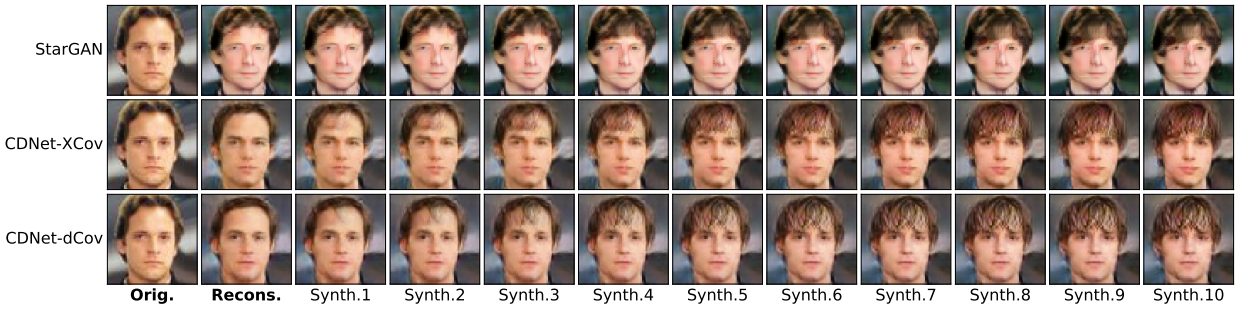
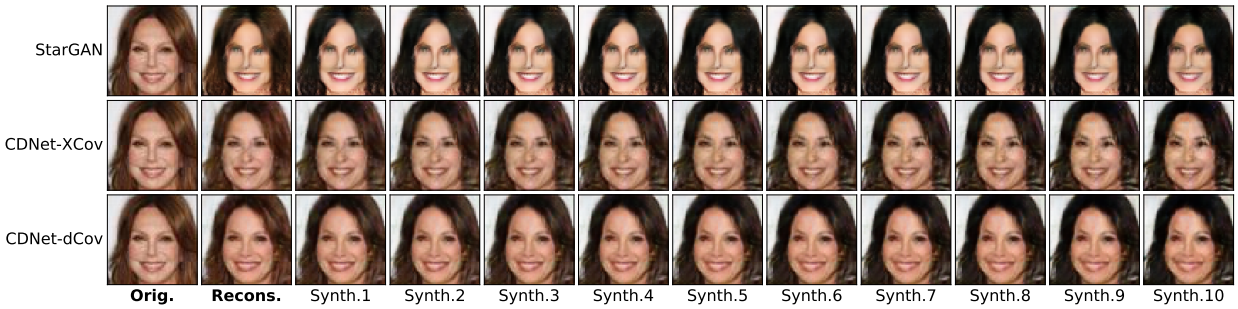


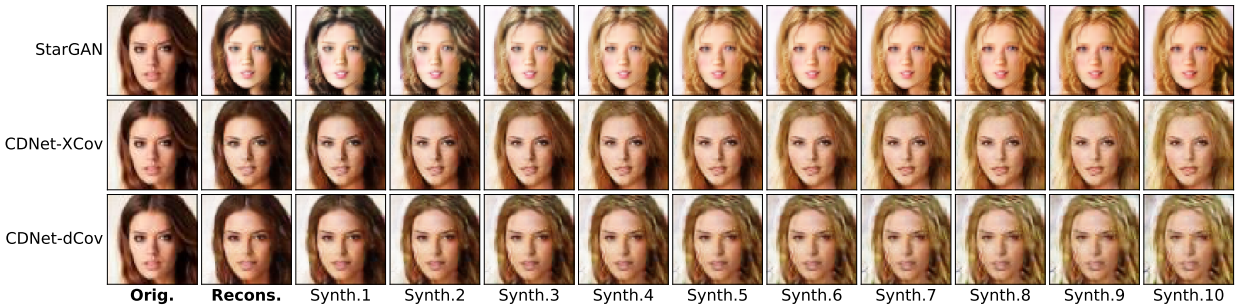
Figure 9: Synthesized face images with the designated attributes. Best viewed in color.



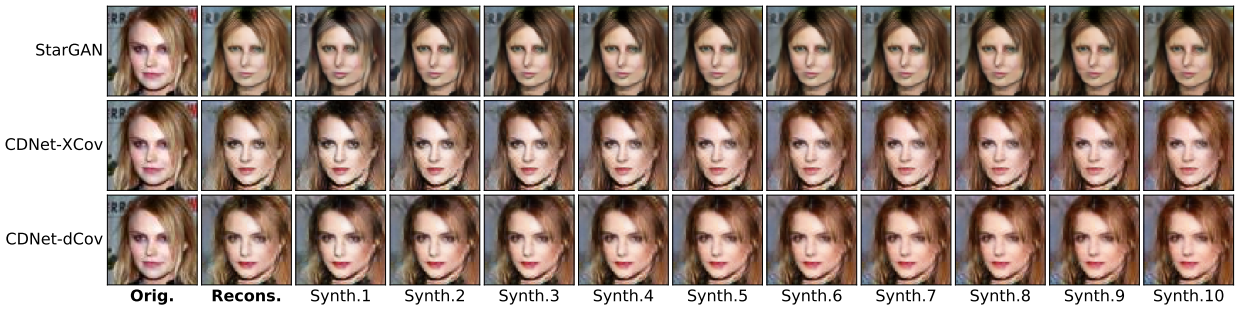
(a) Bangs



(b) Black hair

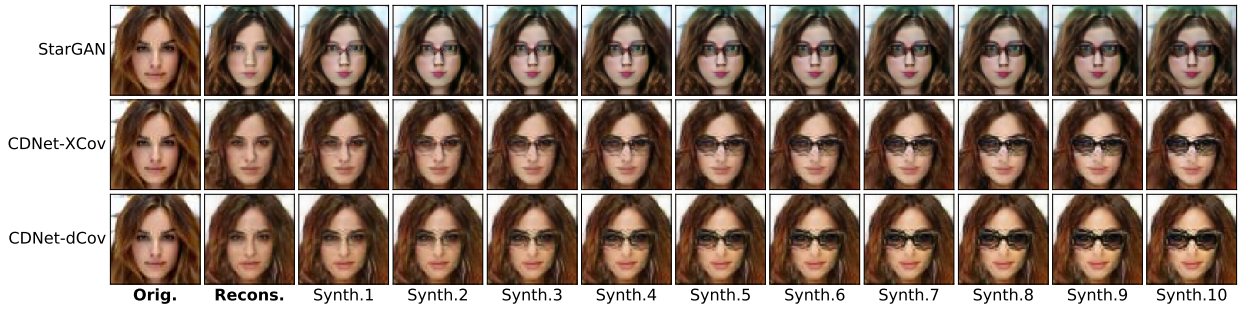


(c) Blond hair

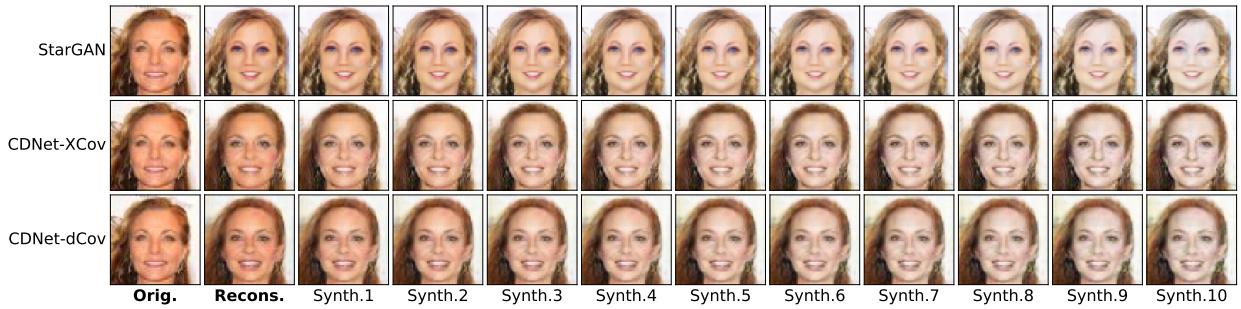


(d) Brown hair

Figure 10: Qualitative comparison of the ability to control degree of disentanglement (Part-1). In each panel, the first column shows the original test image, the second column for reconstructions, and the remaining ten columns for synthesized images with different attribute intensities, from weaker levels to stronger ones. Best viewed in color.

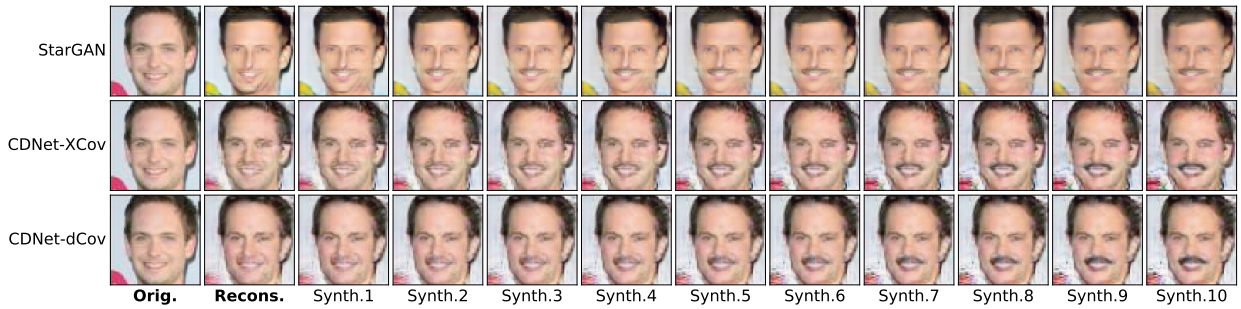


(a) Eyeglasses



(b) Pale skin

(c) Smiling



(d) Mustache

Figure 11: Qualitative comparison of the ability to control degree of disentanglement (Part-2). In each panel, the first column shows the original test image, the second column for reconstructions, and the remaining ten columns for synthesized images with different attribute intensities, from weaker levels to stronger ones. Best viewed in color.

In all aforementioned experiments, StarGAN suffers from a common problem, namely the core object identities of most synthetic faces are obviously changed. It is reasonable to presume that if the objects in two images belong to

different categories, the discriminative features of them should also be distinguishable for classification. To quantify this object difference, we employ the Fréchet Inception Distance (FID) score to measure the feature similarity between original test image sets and generated image sets. Specifically, given a target facial attribute, we first select all test images without containing that attribute as the original test set; then the disentanglement model takes as input images from the original test set to generate new faces (the target attribute's intensity is set to 1), resulting in the generated image set; finally, the FID score is computed on these two image sets. As shown in Figure 12, both two CDNet models have lower FID scores than StarGAN on all target attributes except the "Blond hair" attribute. This result demonstrates that the face images generated by CDNet are more likely to share similar features with the original test images, thus providing higher consistency of the object identities.

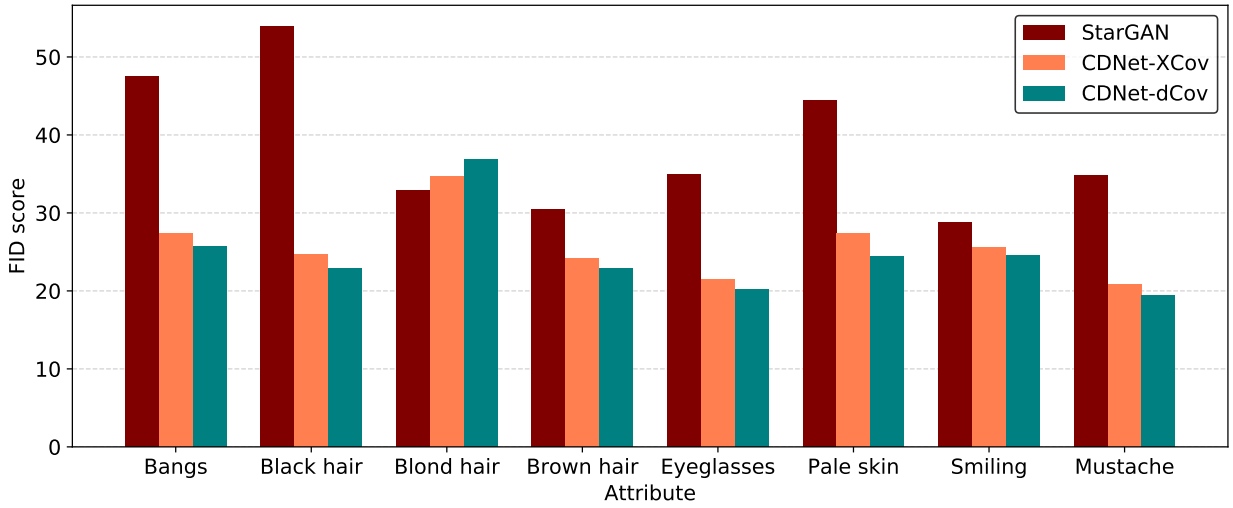


Figure 12: Measurement of feature similarity between original test images and generated images with different facial attributes. The lower the FID score, the more similar the features, and thus the more consistent the object identities.