Supplementary Material: Towards A Controllable Disentanglement Network

Zengjie Song[1], Oluwasanmi Koyejo[2], Jiangshe Zhang[1]
[1]School of Mathematics and Statistics, XJTU, Xi'an, Shaanxi 710049, China
[2]Department of Computer Science, UIUC, Urbana, Illinois 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 (i.e., 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 Ability

The results are shown in Figures 1-5. Best viewed in color.

2.2 Controllable Disentanglement

The results are shown in Figures 6 and 7. Best viewed in color.

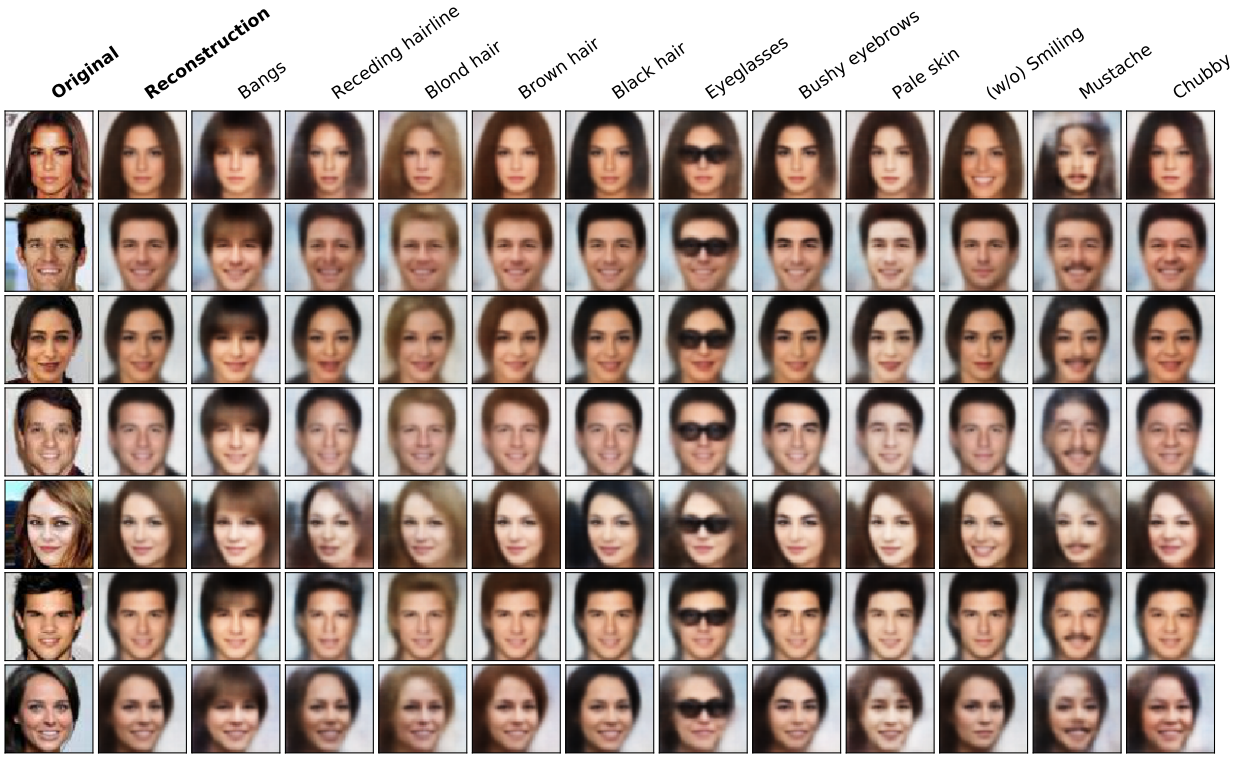


Figure 1: Synthesized face images with the designated attributes by the **AE-XCov** model.

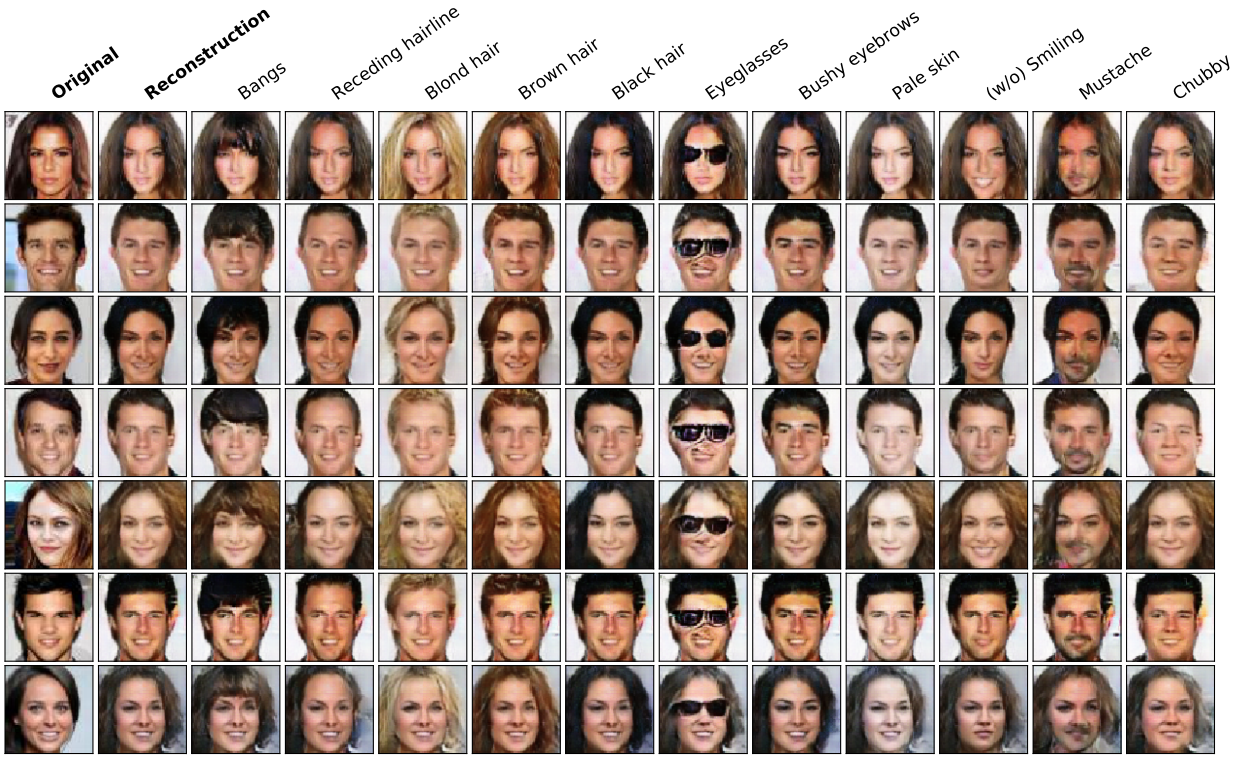


Figure 2: Synthesized face images with the designated attributes by the **IcGAN** model.

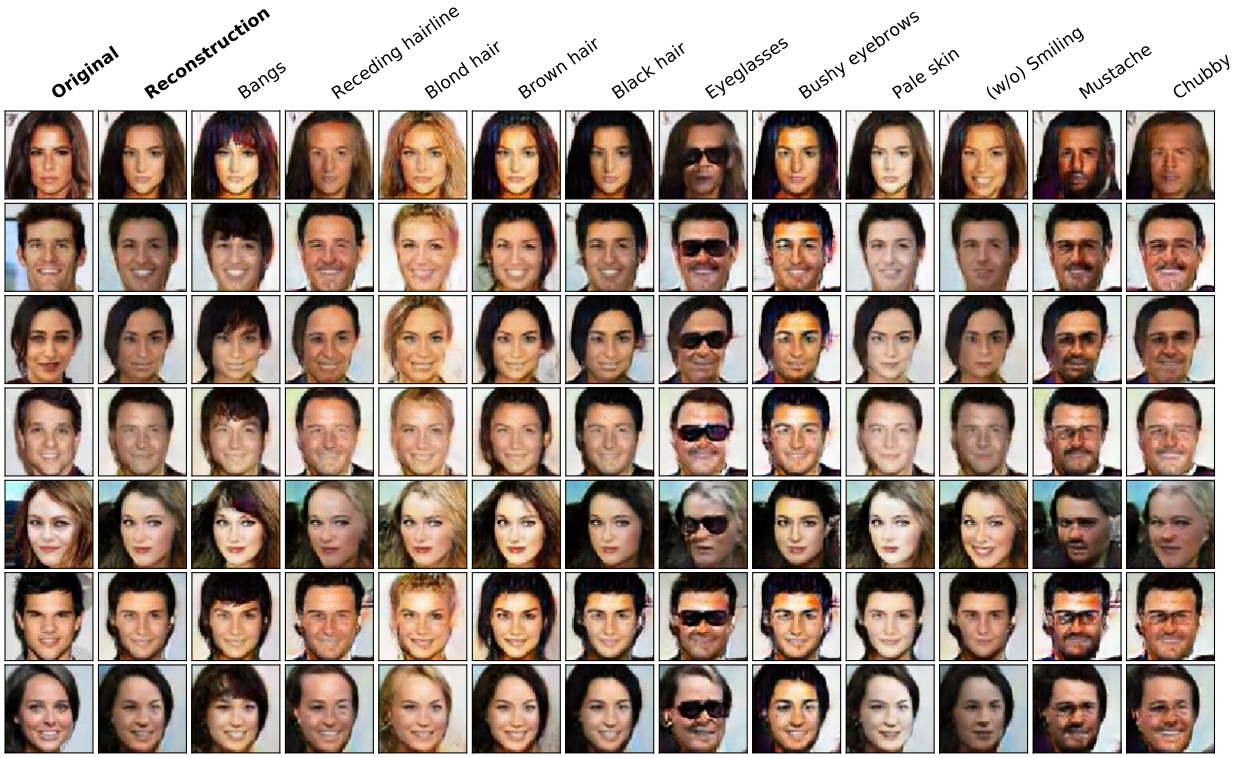


Figure 3: Synthesized face images with the designated attributes by the **VAE/GAN** model.

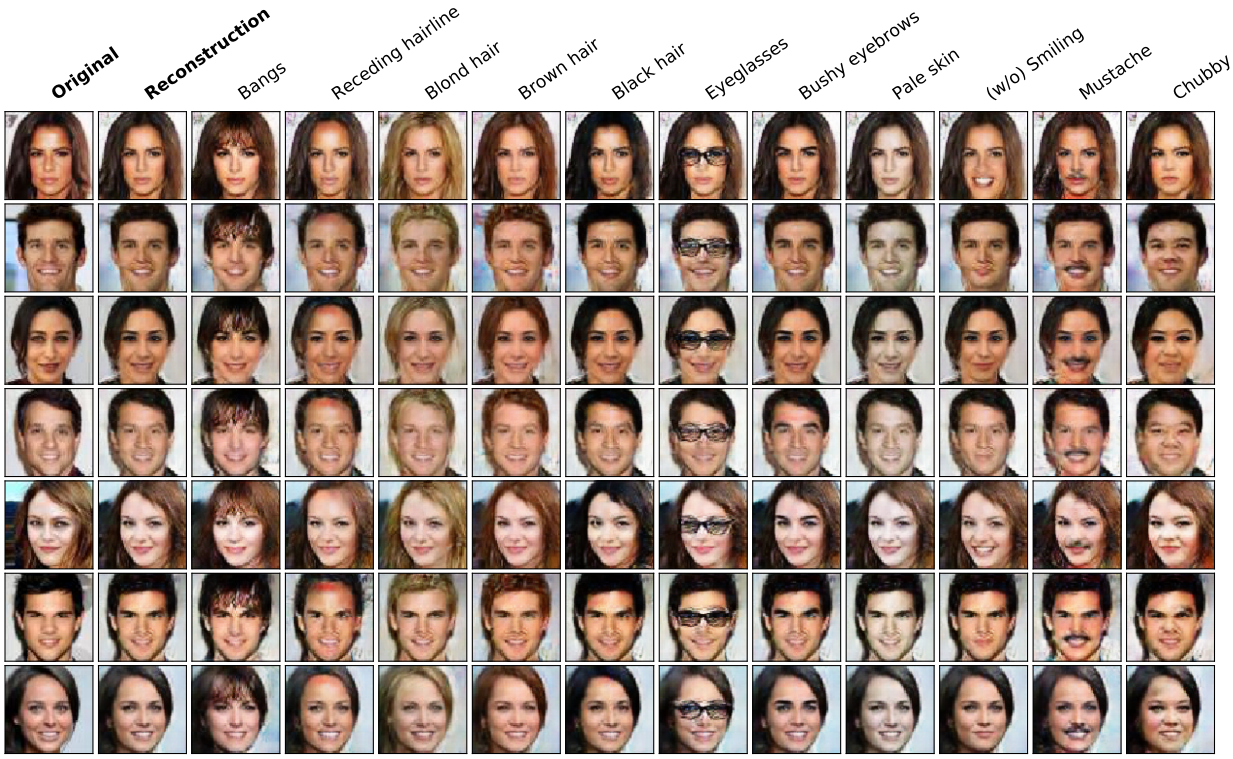


Figure 4: Synthesized face images with the designated attributes by the **CDNet-XCov** model.

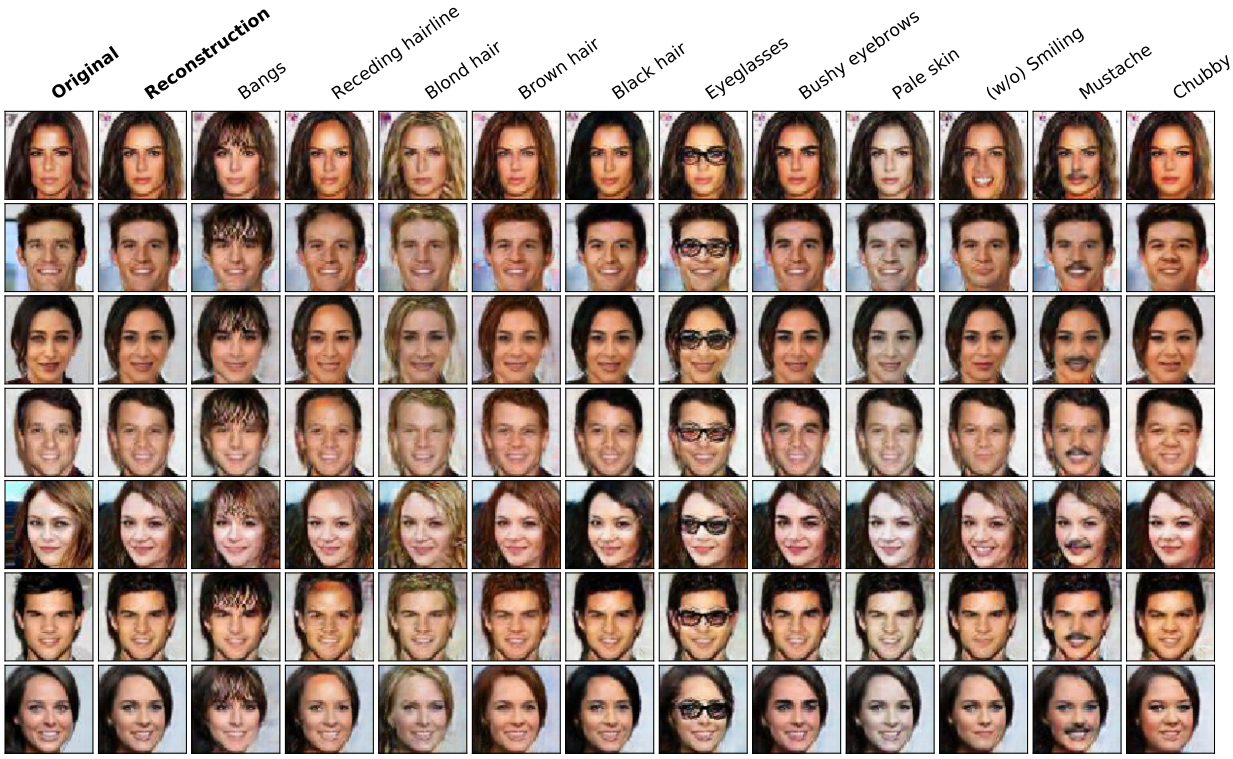
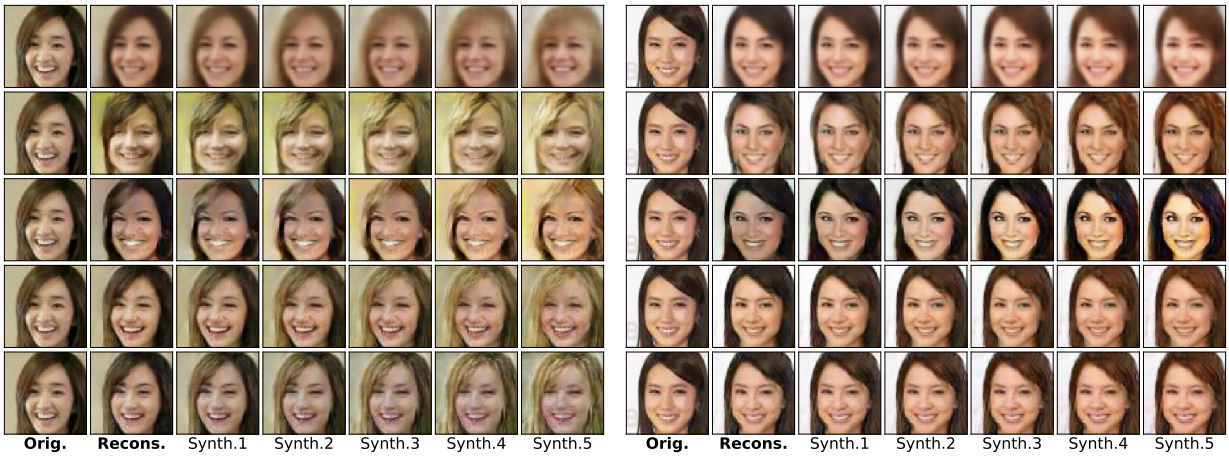


Figure 5: Synthesized face images with the designated attributes by the **CDNet-dCov** model.



(a) Blond hair

(b) Brown hair

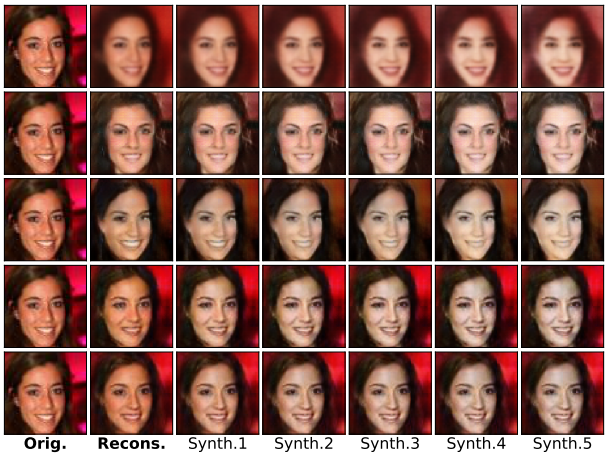
(c) Black hair

(d) Bangs

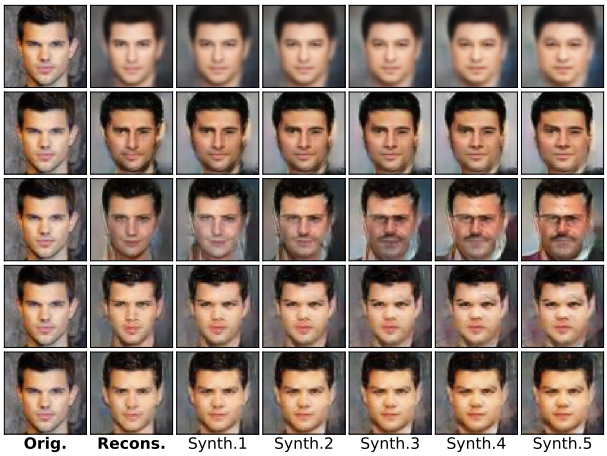
(e) Receding hairline

(f) Mustache

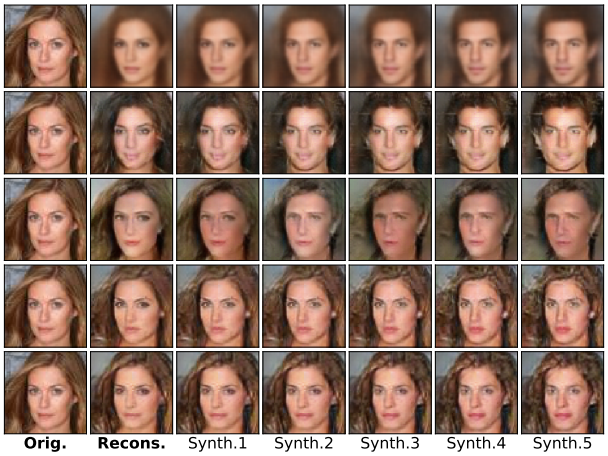
Figure 6: Synthesized face images with different facial attributes and attribute intensities (Part-1). The results in each panel, from the first row to the last row, are obtained by AE-XCov, IcGAN, VAE/GAN, CDNet-XCov, and CDNet-dCov, respectively. In each panel, the first column shows the original test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



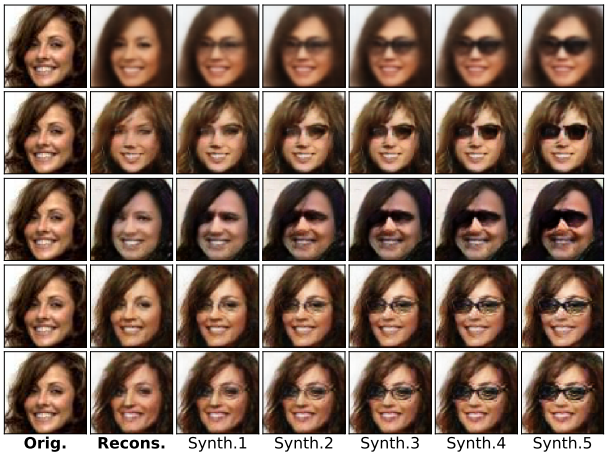
(a) Pale skin



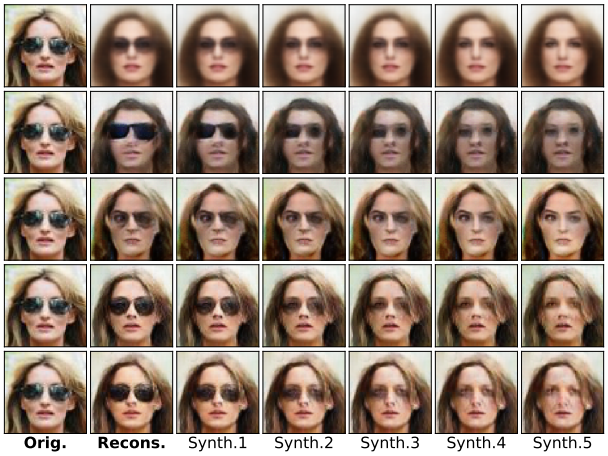
(b) Chubby



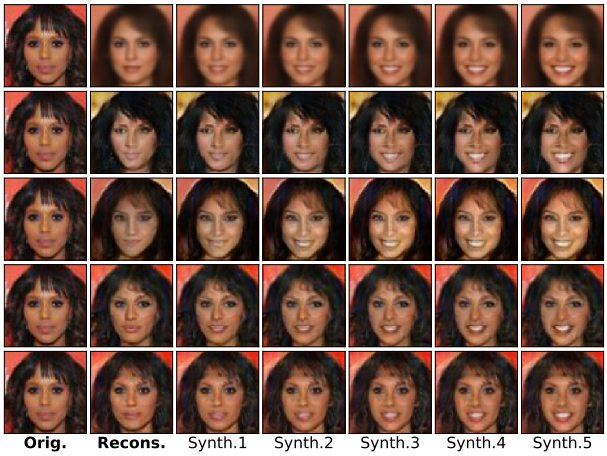
(c) Male



(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.