Learning Controllable Disentangled Representations with Decorrelation Regularization (Supplementary Material)

1 Network Architecture Details

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

Operation	**Kernel**	**Stride**	**Padding**	**Filters**	**BN**	**Activation**	**Dropout**
Convolution	4×4	2×2	1×1×1×1	32	Yes	ReLU	No
Convolution	4×4	2×2	1×1×1×1	64	Yes	ReLU	No
Fully connected	-	-	-	1000	No	Linear	Yes ($p = 0.25$)
Fully connected	-	-	-	$\hat{\mathbf{y}}$:10 $\mathbf{z}$:10	No	$\hat{\mathbf{y}}$:Softmax $\mathbf{z}$:Linear	No
Fully connected	-	-	-	1000	No	Linear	No
Fully connected	-	-	-	3136	Yes	Linear	No
Deconvulution	4×4	2×2	1×1×1×1	32	Yes	ReLU	No
Deconvulution	4×4	2×2	1×1×1×1	1	No	Sigmoid	No

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

Operation	**Kernel**	**Stride**	**Padding**	**Filters**	**BN**	**Activation**	**Dropout**
Convolution	4×4	2×2	1×1×1×1	16	Yes	ReLU	No
Convolution	4×4	2×2	1×1×1×1	32	Yes	ReLU	No
Convolution	4×4	2×2	1×1×1×1	64	Yes	ReLU	No
Fully connected	-	-	-	4000	No	Linear	Yes ($p = 0.5$)
Fully connected	-	-	-	2000	No	Linear	Yes ($p = 0.5$)
Fully connected	-	-	-	$\hat{\mathbf{y}}$:40 $\mathbf{z}$:1000	No	$\hat{\mathbf{y}}$:Sigmoid $\mathbf{z}$:Linear	No
Fully connected	-	-	-	2000	No	Linear	No
Fully connected	-	-	-	4000	No	Linear	No
Fully connected	-	-	-	4096	Yes	Linear	No
Deconvulution	4×4	2×2	1×1×1×1	32	Yes	ReLU	No
Deconvulution	4×4	2×2	1×1×1×1	16	Yes	ReLU	No
Deconvulution	4×4	2×2	1×1×1×1	3	No	Tanh	No

2 Additional Results

Here, we first provide a bigger version of the Figure 2 in our paper to illustrate the image manipulation methods, as shown in the following Figure 1. Then, we give more synthesized face images based on the CelebA test set.

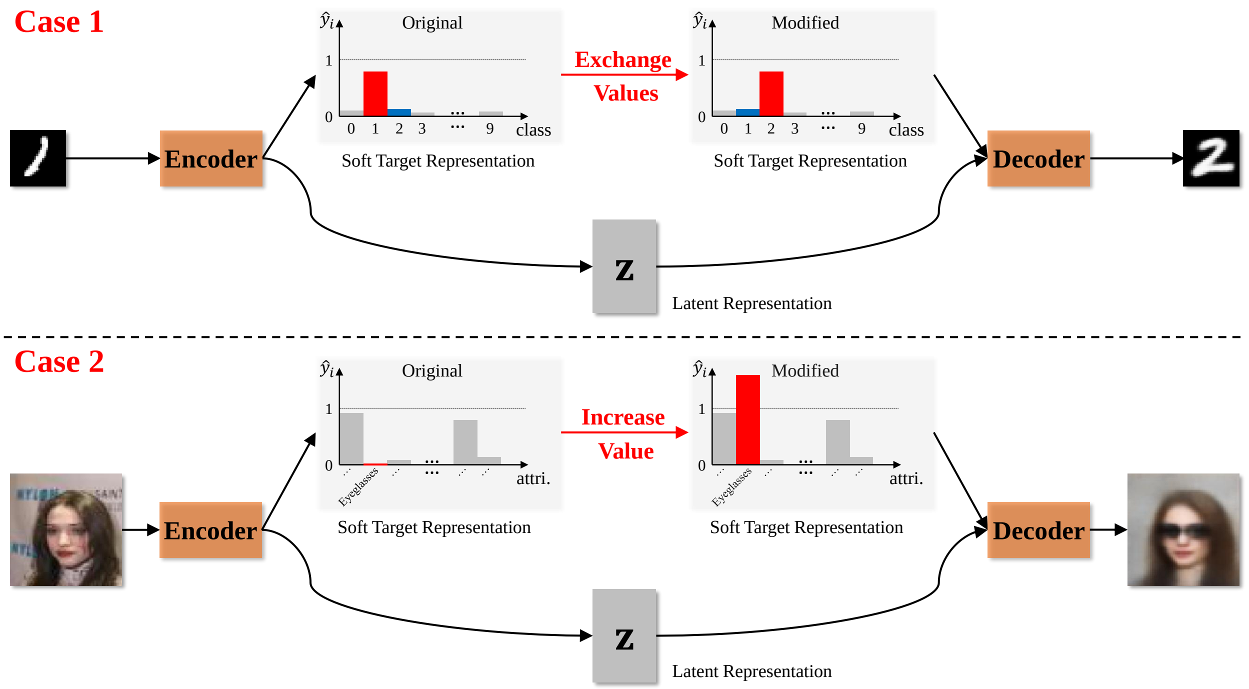


Figure 1: Manipulating images. **Case 1**: for the disentanglement task where the mutual exclusion exists among multiple classes. Specifically, we first employ the encoder to infer the soft target representation $\hat{\mathbf{y}}$ and the latent representation $\mathbf{z}$ of the digit "1" in boldface. Then we modify $\hat{\mathbf{y}}$ by exchanging the third element (corresponding to digit 2 class) and the maximum element (ideally corresponding to digit 1 class), while keeping remaining elements fixed. Finally, we feed the modified $\hat{\mathbf{y}}$ and the unchanged $\mathbf{z}$ to the decoder to generate the new digit "2" which is also in boldface. **Case 2**: for the scenario where multiple attributes are independent of each other. The overall procedure is similar to the Case 1, but with a different modification of $\hat{\mathbf{y}}$. Specifically, in order to generate a new face with eyeglasses, we just replace the original (near) zero value corresponding to "Eyeglasses" attribute with the new value (e.g., 1.7) in $\hat{\mathbf{y}}$.

2.1 Verifying the Disentanglement Ability

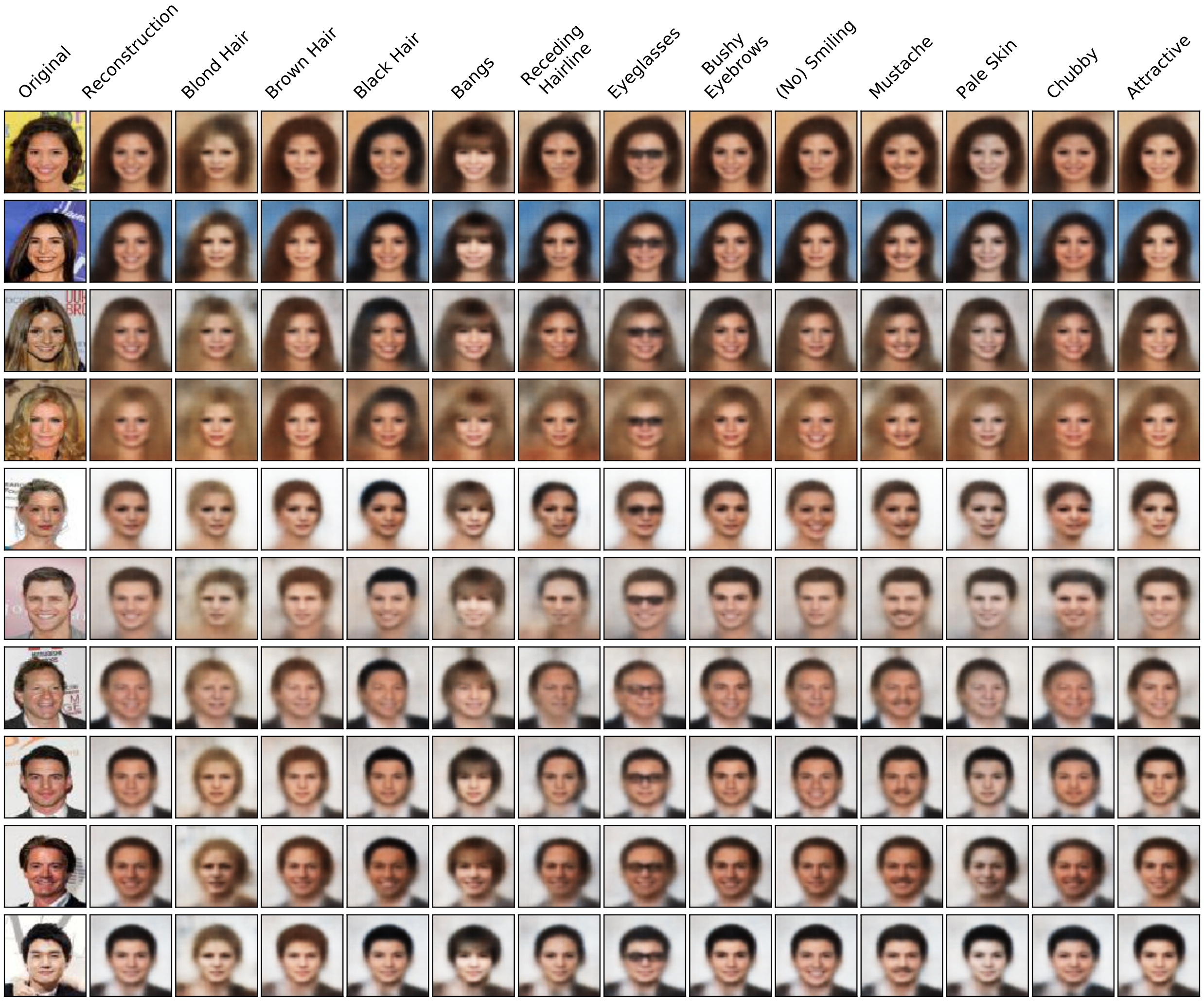


Figure 2: Synthesized face images with the designated attributes by the **disAE-XCov** model.

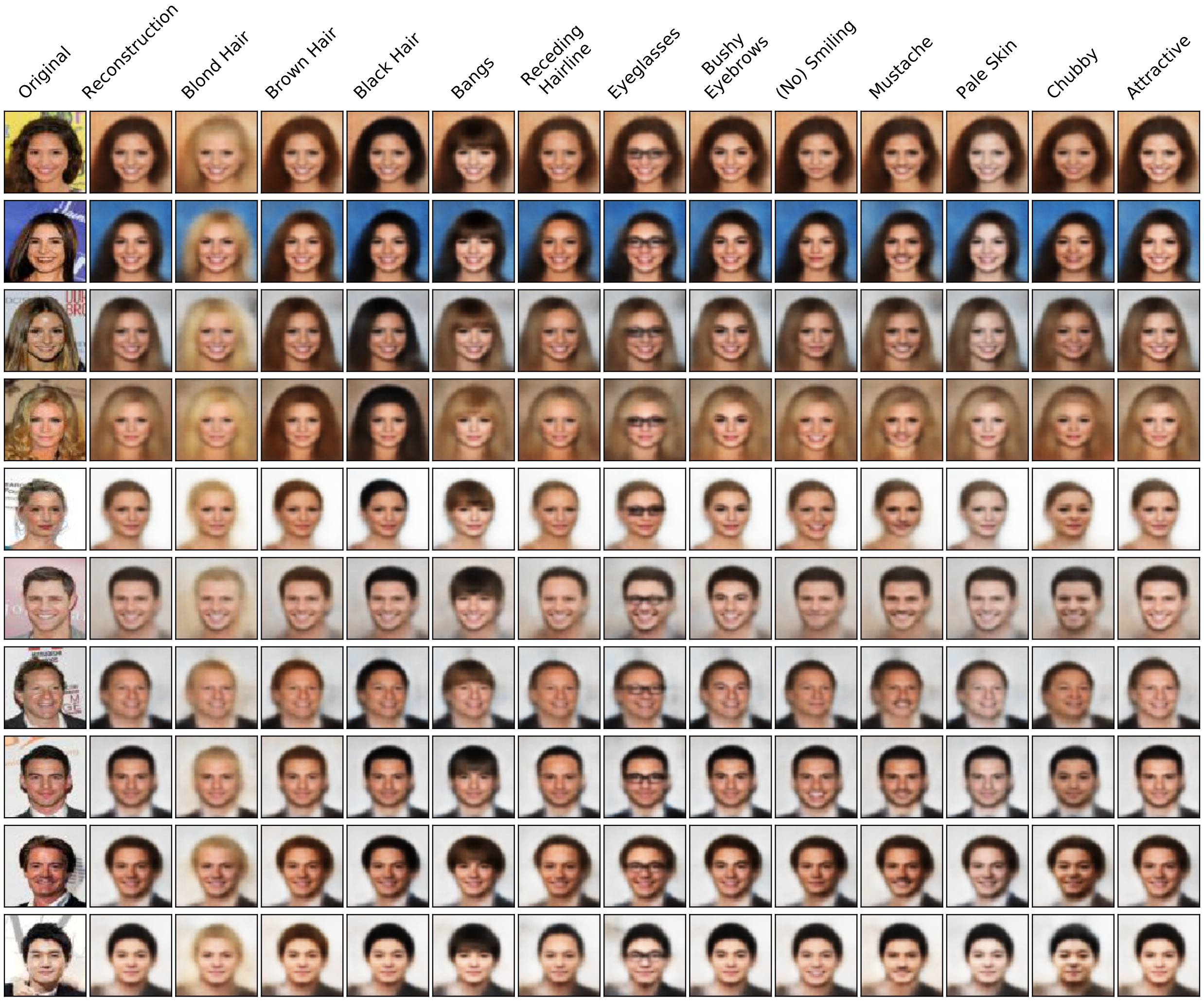


Figure 3: Synthesized face images with the designated attributes by the **mddAE-XCov** model.

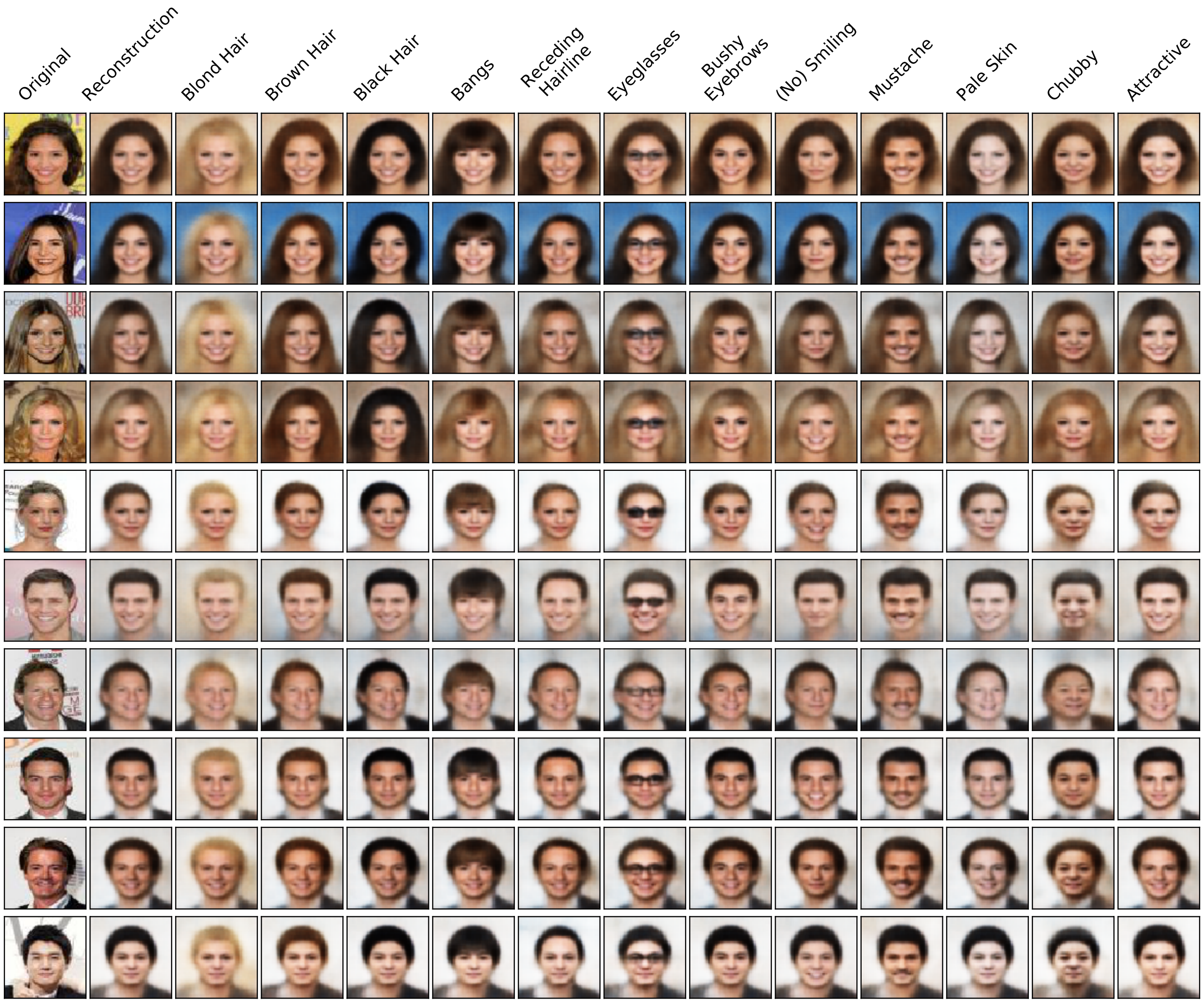
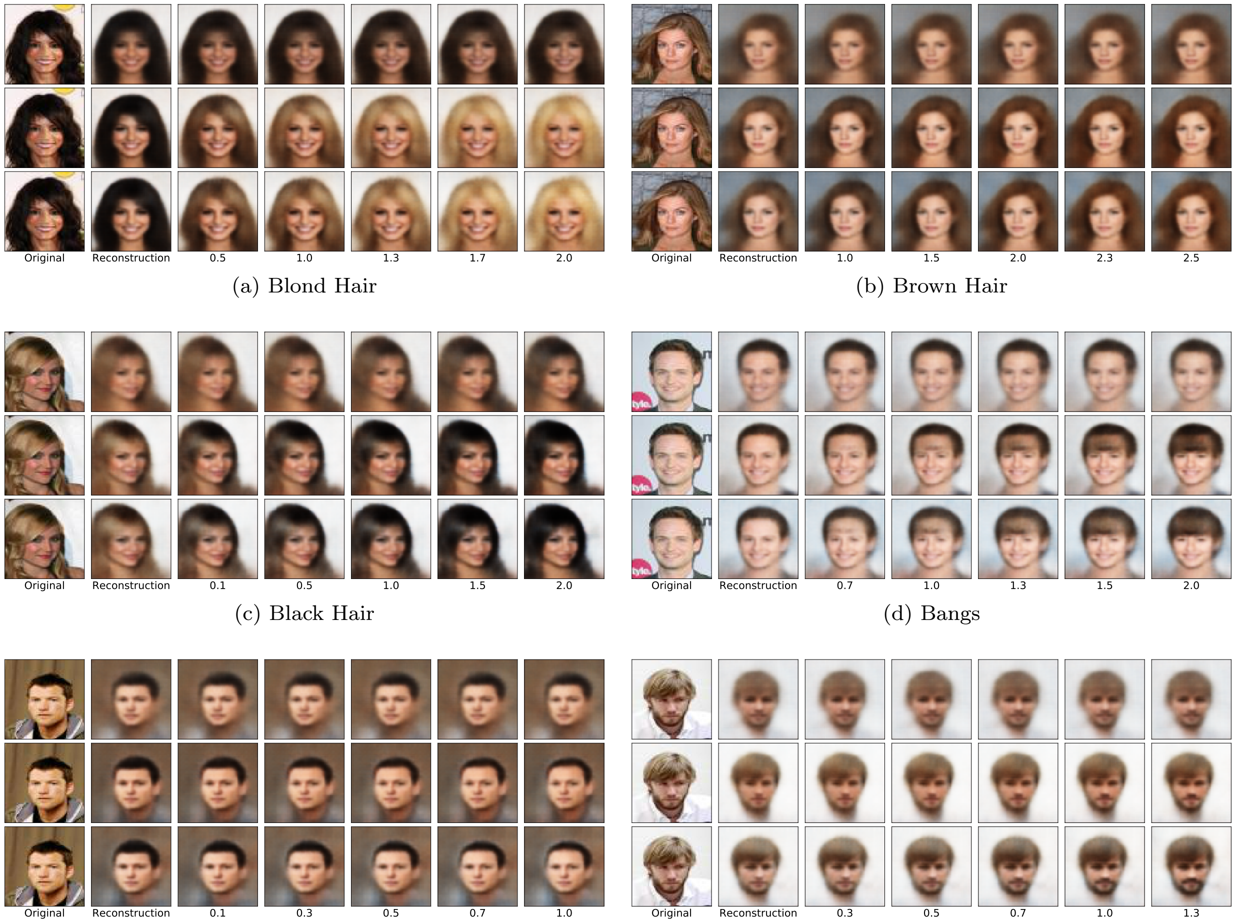


Figure 4: Synthesized face images with the designated attributes by the **mddAE-dCov** model.

2.2 Controllable Disentanglement



(a) Blond Hair

(b) Brown Hair

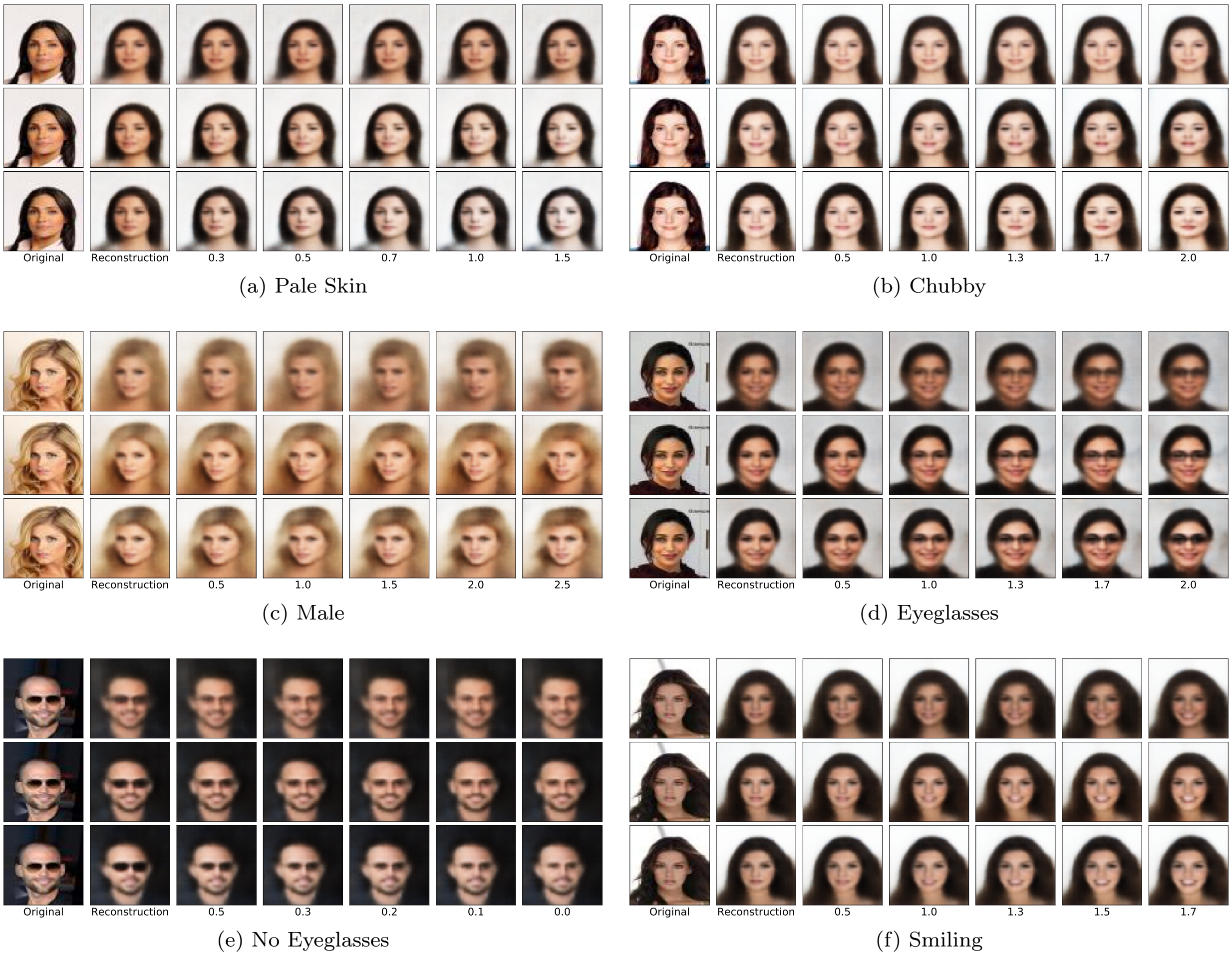
(c) Black Hair

(d) Bangs

(e) Receding Hairline

(f) Mustache

Figure 5: Synthesized face images with different facial attributes and attribute intensities (Part-1). In each panel, the **first row** corresponds to the results of disAE-XCov, the **second row** for mddAE-XCov, the **third row** for mddAE-dCov, and the attribute intensity values are given below the third row.



(a) Pale Skin

(b) Chubby

(c) Male

(d) Eyeglasses

(e) No Eyeglasses

(f) Smiling

Figure 6: Synthesized face images with different facial attributes and attribute intensities (Part-2). In each panel, the **first row** corresponds to the results of disAE-XCov, the **second row** for mddAE-XCov, the **third row** for mddAE-dCov, and the attribute intensity values are given below the third row.

2.3 Comparing Disentanglement Strength by Classification

These experiments are used to quantitatively compare the disentanglement strength of two decorrelation regularizations, namely cross covariance (XCov) and distance covariance (dCov). The specific comparison protocol is illustrated in Section 4.4 of the paper.

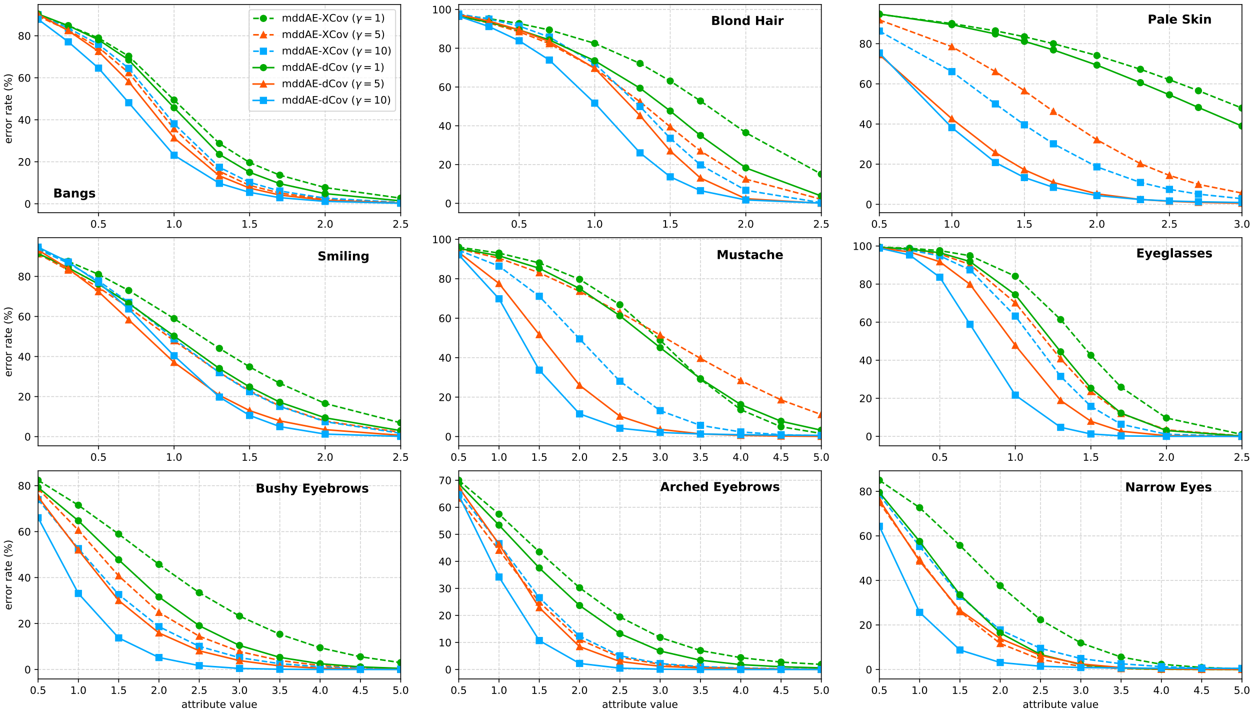


Figure 7: The attribute classification of synthesized face images by mddAE-XCov and mddAE-dCov (Part-1). For each attribute, a linear SVM is ran for 5 times and we report the average performance.

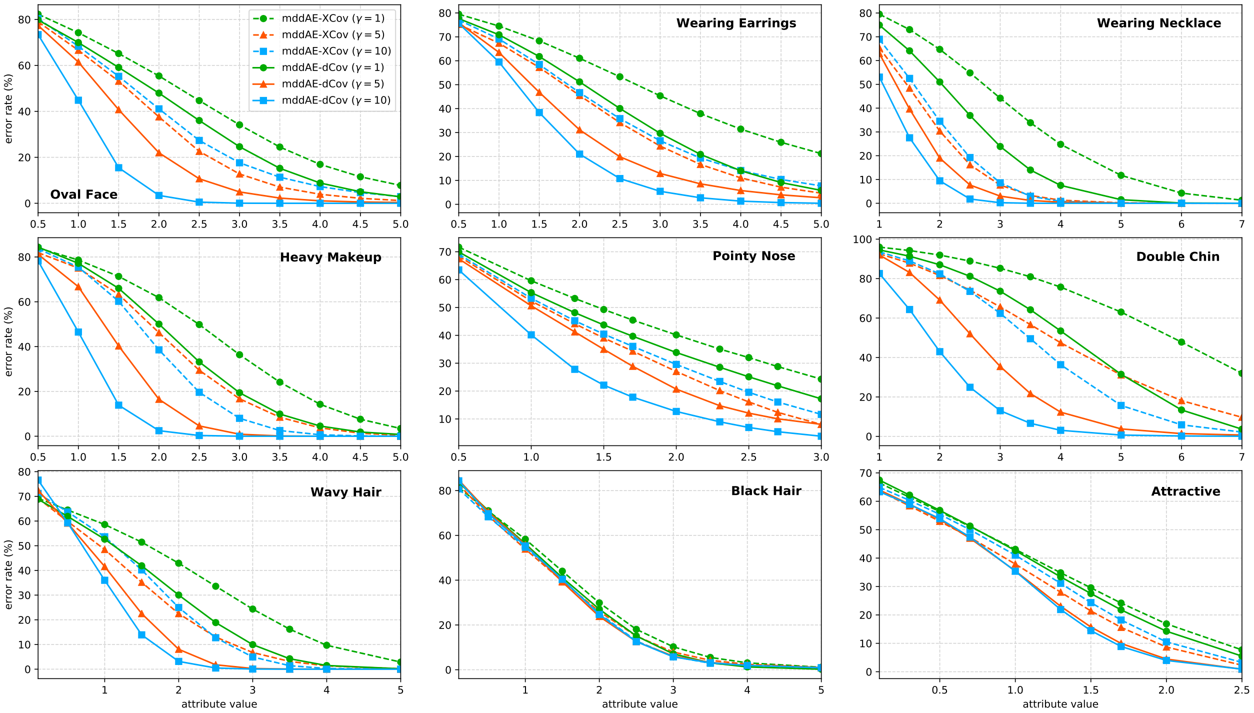


Figure 8: The attribute classification of synthesized face images by mddAE-XCov and mddAE-dCov (Part-2). For each attribute, a linear SVM is ran for 5 times and we report the average performance.