Supplementary Material for Cross-GAN Auditing: Unsupervised Identification of Attribute Level Similarities and Differences between Pretrained Generative Models

Matthew L. Olson[1], Shusen Liu[2], Rushil Anirudh[2], Jayaraman J. Thiagarajan[2], Peer-Timo Bremer[2], and Weng-Keen Wong[1]
[1] Oregon State University - EECS, [2]Lawrence Livermore National Laboratory - CASC
{olsomatt,wongwe}@oregonstate.edu, {liu42,anirudh1,jayaramanthi1,bremer5}@llnl.gov

A. Ablation study

First, we investigate the effect of the λ_b parameter on KLIEP loss that allows us to discover novel attributes. In addition to KLIEP loss presented in the main text, we analyze a model trained with simple log loss used to train binary classifiers to predict the likelihood of a given sample.

A.1. Log Loss Model

This model, we denote as LOG, is nearly identical to the DRE models except instead of a softplus final activation, it uses a sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$. These models are used to classify whether a given feature belongs to $\mathcal{G}_c(z)$ or $\mathcal{G}_r(\bar{z})$. We pre-train two separate LOG models to approximate $\hat{\gamma}_c(x) = \hat{p}_c(x)$, and $\hat{\gamma}_r(x) = \hat{p}_r(x)$, where we treat $\mathcal{G}_c$ as $P(\mathrm{x}|Y=1)$ and $\mathcal{G}_r$ as $P(\mathrm{x}|Y=0)$. These LOG models are learned using simple 2-layer MLPs, $f^c_{LOG}(\,), f^r_{LOG}(\,)$, such that

$$\hat{\gamma}_c(\mathrm{z}) = f^c_{\mathrm{LOG}}(\mathcal{F}(\mathcal{G}_c(\mathrm{z}))) \text{ and } \hat{\gamma}_r(\bar{\mathrm{z}}) = f^r_{\mathrm{LOG}}(\mathcal{F}(\mathcal{G}_r(\bar{\mathrm{z}}))), \tag{1}$$

where $\mathcal{F}$ is the same Encoder model used in main paper's equation 3.

The loss used for training the LOG models is defined as follows:

$$\mathcal{L}^c_{\mathrm{Log}} = \frac{1}{T_2}\sum_{j=1}^{T_2} -\log(1-\hat{\gamma}_c(\bar{\mathrm{z}}_j)) + \frac{1}{T_1}\sum_{i=1}^{T_1} -\log(\hat{\gamma}_c(\mathrm{z}_i)) \tag{2}$$

where $\bar{\mathrm{z}}_j$ and z_i are random samples drawn from the latent space of each generator. The loss term for the second model LOG model is

$$\mathcal{L}^r_{\mathrm{Log}} = \frac{1}{T_1}\sum_{j=1}^{T_1} -\log(1-\hat{\gamma}_r(\bar{\mathrm{z}}_j)) + \frac{1}{T_2}\sum_{i=1}^{T_2} -\log(\hat{\gamma}_r(\mathrm{z}_i)) \tag{3}$$

The LOG models $f^1_{LOG}(\,), f^2_{LOG}(\,)$ are trained to minimize $\mathcal{L}^1_{\mathrm{Log}}, \mathcal{L}^2_{\mathrm{Log}}$ respectively.

λ	$\mathcal{R}_{\mathrm{Score}}$ (DRE loss) (↑)	$\mathcal{R}_{\mathrm{Score}}$ (Log loss)(↑)
0	0.42 ± 0.38	0.42 ± 0.38
0.1	$\mathbf{0.61} \pm 0.35$	0.37 ± 0.41
0.2	0.54 ± 0.33	0.44 ± 0.39
0.5	0.57 ± 0.40	0.45 ± 0.38
1	$\mathbf{0.61} \pm 0.33$	0.40 ± 0.40
5	0.57 ± 0.39	0.34 ± 0.32

Table 1. The effect on the unique direction score when modifying the regularization λ on the average $\mathcal{R}_{\mathrm{Score}}$ ($\pm$ std) for the the 7 CelebA pairwise leave-attribute-out experiments using a Robust ResNet-50 encoder.

Finally, the trained LOG models are used to minimize the loss in equation 7 (rather than DRE models); the objective in equation 7 remains the same.

A.2. Missing attribute ablation study results

Table 1 illustrates the missing attribute discovery score for each CelebA split versus full CelebA. With $\lambda = 0$ (i.e. ignoring the DRE loss), the attribute discovery process has difficulty capturing some missing attributes. When using a regularization model trained with Log-loss, the results are consistently worse than DRE, sometimes even worse than with $\lambda = 0$. The KLIEP loss model, on the other hand, performs consistently better for all lambda values > 0.

B. Same dataset, different architecture

To verify the effectiveness of xGA at comparing models trained on the same dataset with different configurations, we perform two sets of experiments. We use Prog-GAN [2] (client) and GANformer [1] (reference) trained on the FFHQ dataset. Figure 1 shows an example of how these two GANs can be aligned, and how the novel/missing attribute reflects each GAN's capacity to learn the data distribution. We also use two configurations of a StyleGAN3 [3] trained on FFHQ. Figure 2 shows how translation equivarience is

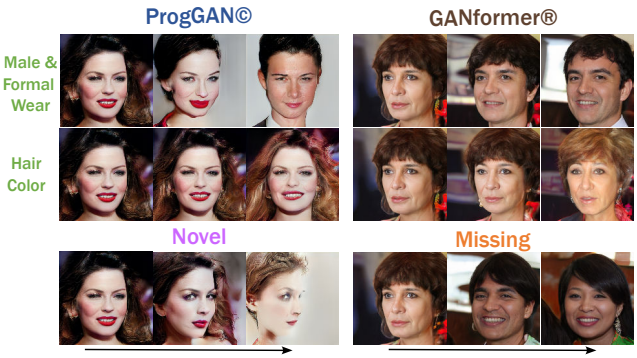

Figure 1. An example of applying our method to two generative models trained on the same dataset (FFHQ). We find ProgGAN and GANformer are able to find some alignment, and that the newer model (GANformer) is better at capturing the full data distribution of FFHQ (Missing) whereas ProgGAN is prone to generating non-realistic images (Novel).

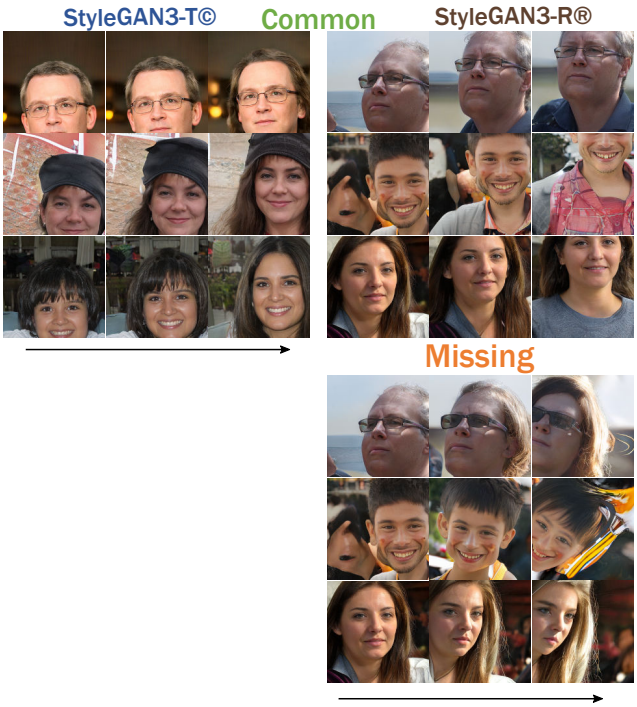

Figure 2. A few examples from our experiment applying xGA between two StyleGAN3 models, both trained on FFHQ, but with different model configurations. As expected both models having translation equivarience, and the rotation equivariance is missing from the translation model.

preserved in both models, whereas only the StyleGAN3-r model is rotationally equivarient.

C. Extending xGA to compare multiple GANs

Though all our experiments used a client model w.r.t a reference, our method can be readily extended to perform

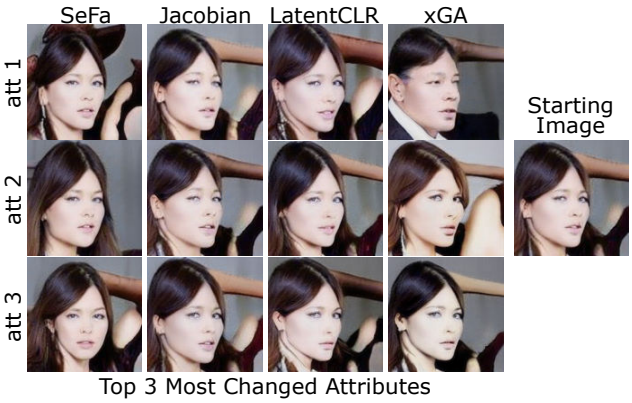

Figure 3. Comparing xGA on single GAN attribute discovery with existing approaches, we find that more diverse and novel attributes can be found simply by using an external feature space. We exploit this for effective alignment across two GAN models. Complete examples for all methods are provided below.

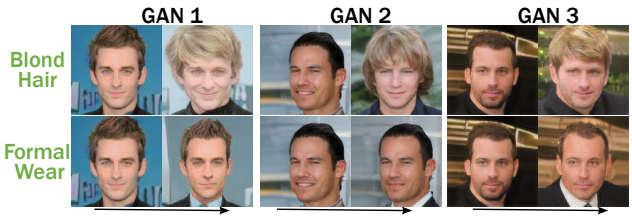

Figure 4. Common attributes identified using xGA with three different StyleGANs.

comparative analysis of multiple GANs, with the only constraint arising from GPU memory since all generators need to be loaded into memory for optimization. We performed a proof-of-concept experiment by discovering common attributes across 3 different independently trained StyleGANs as shown in Figure 4. For this setup, we expanded the cost function outlined in equation 3 to include 3 pairwise alignment terms from the 3 GANs to perform contrastive training, in addition to an extra independent term from the third model. While beyond scope for the current work, scaling xGA is an important direction for future work.

D. Single GAN results

Here we present the full training of all learned directions for each of our methods using the same starting point from CelebA GAN. Figure 3 visualizes a shortened example of the top 3 attributes (induce most changes in the "oracle" classifier predictions) and it is clear that xGA identifies the most diverse semantic changes. Complete results can be seen as follows:

1. Sefa [5]: Figure 15

2. LatentCLR [8]: Figure 13

3. Voynov [6]: Figure 14
4. Hessian [4]: Figure 11
5. Jacobian [7]: Figure 12
6. xGA (ImageNet ResNet-50): figure 6
7. xGA (advBN ResNet-50): figure 7
8. xGA (CLIP ResNet-50): figure 8

We visualize both positive and negative directions for every model. Even though xGA and LatentCLR are not directly trained for negative directions, we find these attributes to be semantically meaningful and interesting.

Next we present an example where we compare two LatentCLR models trained on different GANs where the reference GAN is CelebA and client GAN is CelebA without Hats. We sort all the directions by most similar (as described in the main paper) and show an example of the results in Figure 5, finding many similarities, but no dedicated Hat attribute in the reference GAN. Showing how without the dedicated constraint of the DRE models, finding missing attributes is difficult.

E. Expanded Qualitative Results

Here we present many additional examples of shared directions between two GANs, and novel/missing directions from a few different GAN pairs that contain subset of the CelebA dataset. We introduce a new GAN (anime), as it produces interesting common, missing, and novel attributes, though the GAN itself produces lower quality images than other models, and as such we leave it here in the supplement. The figures are arranged as follows:

1. Common attributes: CelebA (reference) and Metface (client) sketch (16), formal (17), and curly hair (18)
2. Common attributes: Anime (client) and Toon (reference) purple hair (19), orange/brown hair (20), open mouths (21), and smiling (22); missing attributes of green hair / lipstick (23)
3. Common attributes: CelebA (reference) and Disney (client) blonde hair (24), and brown hair (25); novel Disney attributes of turning green / cartoonish eyes (26)
4. Additional missing attributes from different CelebA client GANs, with CelebA reference GAN (27)

F. Expanded Quantitative Results

First we present the results for using ViT-based feature extractors in table 4. We include 3 different pretrained models: one original trained on ImageNet, CLIP, and MAE. While ViT does well for entropy metric, it performs poorly for cross model based experiments.

Next, we present the entire results for our missing attribute quantitative experiments. To recap these experiments, we use the 7 controlled CelebA models which are missing one or more attributes (hat, glasses, male, female, beard, beards—hats, and smiles—glasses—ties) and treat them as the client model; we audit these models with respect to the reference CelebA GAN. The 7 missing attribute experiments are shown in table 2, where we can see xGA performs well (e.g., easily finding the missing glasses attribute). The 7 attribute alignment experiments are shown in table 3, where again we see xGA with a robust resnet performs well, especially when the client GAN is missing multiple attributes (e.g., client GAN is missing beards and hats).

For completion's sake, we run pairwise experiments between each GAN, treating each GAN as reference versus the other 7 GANs, which results in a total of 56 client/reference paired experiments. We report these comprehensive results in the following tables (where rows are reference GAN and columns are the client): 9, 8, 7, 6, 10 , 11 , 12 , 13 , 14 , 15 , and 16. We also compute the the common attribute results experiments in the following tables: 20,19,18,21, 22, 23, 24, 25, 26, 27, and 28.

	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
SeFa	0.143	0.143	0.045	0.111	0.063	0.278	0.189
jacobian	0.478	0.536	0.086	0.390	0.388	0.120	0.287
Hessian	1.000	1.000	0.056	0.167	0.167	0.096	0.407
LatentCLR	1.000	1.000	0.250	0.333	0.200	0.153	0.537
Voynov	0.500	1.000	0.333	0.050	0.500	0.153	0.259
xGA (ResNet-50)	1.000	1.000	0.333	0.500	0.200	0.167	0.465
xGA (Clip ResNet-50)	1.000	1.000	1.000	0.200	0.200	0.108	0.383
xGA (advBN ResNet-50)	1.000	1.000	0.250	1.000	0.063	0.183	0.401

Table 2. The full results for the recovery scores ($\mathcal{R}_{\text{score}}$), where CelebA GAN is the reference.

	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
SeFa	0.413	0.458	0.372	0.374	0.314	0.387	0.355
Hessian	0.475	0.489	0.652	0.598	0.618	0.615	0.525
LatentCLR	0.519	0.511	0.556	0.533	0.512	0.593	0.579
Voynov	0.566	0.477	0.567	0.555	0.570	0.562	0.513
Jacobian	0.523	0.452	0.505	0.528	0.519	0.491	0.495
xGA (ResNet-50)	0.457	0.403	0.740	0.792	0.461	0.643	0.489
xGA (Clip ResNet-50)	0.753	0.451	0.772	0.791	0.894	0.580	0.656
xGA (advBN ResNet-50)	0.615	0.357	0.750	0.825	0.619	0.803	0.649

Table 3. The full results for the alignment scores ($\mathcal{A}_{\text{score}}$), where CelebA GAN is the reference

Method / Model	$\mathcal{H}_{\text{score}}$ ($\downarrow$)	$\mathcal{A}_{\text{score}}$ ($\uparrow$)	$\mathcal{R}_{\text{score}}$ ($\uparrow$)
xGA + ViT	1.988 ± 0.068	0.377 ± 0.090	0.249 ± 0.217
xGA + ViT + MAE	2.102 ± 0.035	0.349 ± 0.089	0.194 ± 0.197
xGA + ViT + Clip	2.091 ± 0.041	0.397 ± 0.122	0.268 ± 0.195

Table 4. ViT-based extractors results. The average entropy scores for all 8 CelebA experiments, the average alignment scores ($\mathcal{A}_{\text{score}}$) for the CelebA pairwise experiments, and the average recovery scores ($\mathcal{R}_{\text{score}}$) for the CelebA pairwise leave-attribute-out experiments ($\pm$ std)

References

[1] Drew A Hudson and C. Lawrence Zitnick. Compositional transformers for scene generation. *Advances in Neural Information Processing Systems NeurIPS 2021*, 2021. 1

[2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 1

[3] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[4] William Peebles, John Peebles, Jun-Yan Zhu, Alexei Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *European Conference on Computer Vision*, pages 581–597. Springer, 2020. 3

[5] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1532–1540, 2021. 2

[6] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning*, pages 9786–9796. PMLR, 2020. 3

[7] Yuxiang Wei, Yupeng Shi, Xiao Liu, Zhilong Ji, Yuan Gao, Zhongqin Wu, and Wangmeng Zuo. Orthogonal jacobian regularization for unsupervised disentanglement in image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6721–6730, 2021. 3

[8] Oğuz Kaan Yüksel, Enis Simsar, Ezgi Gülperi Er, and Pinar Yanardag. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14263–14272, 2021. 2

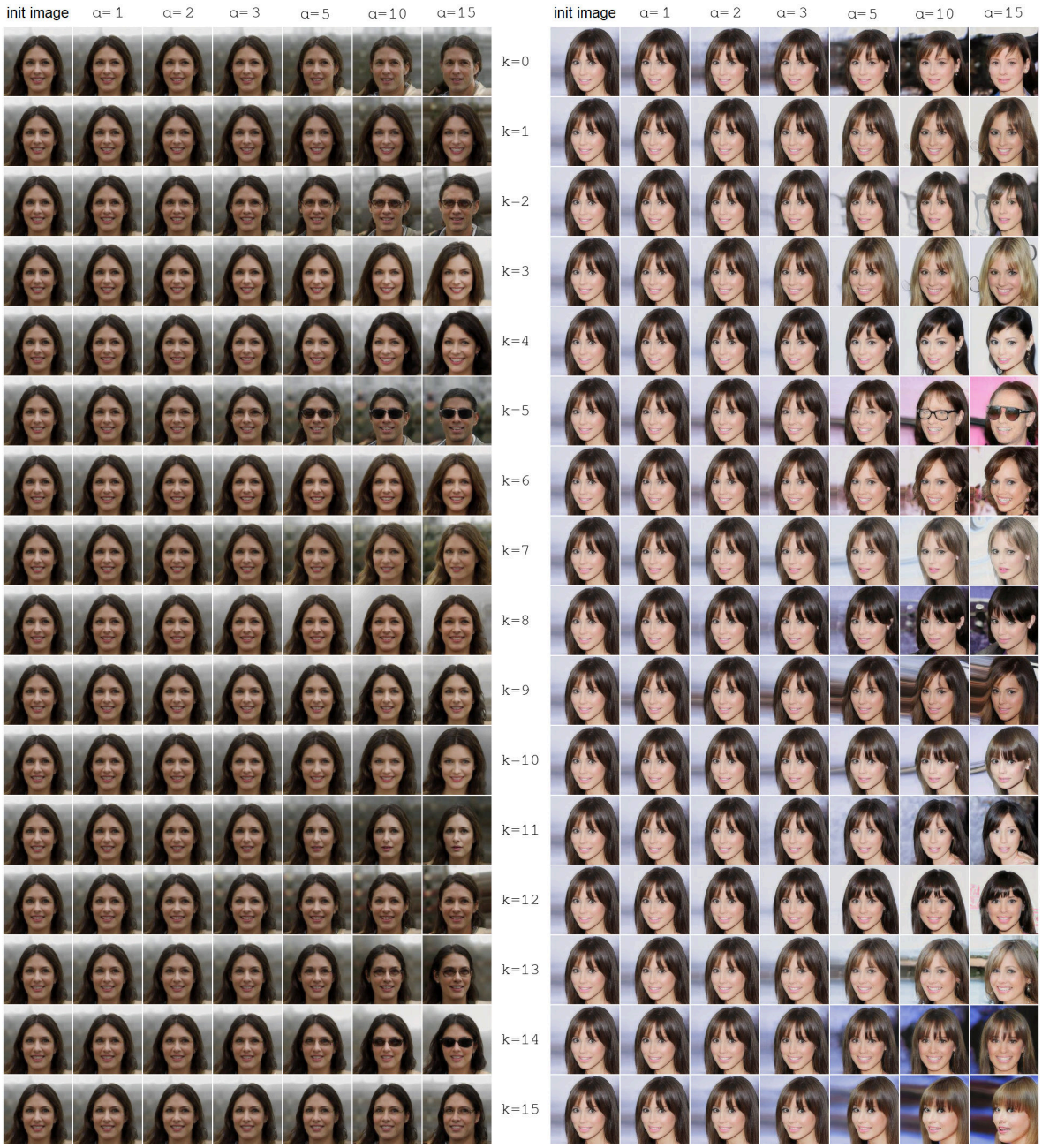

Figure 5. All 16 directions of two single GAN LatentCLR models trained on different GANs where the reference GAN is CelebA and client GAN is CelebA without Hats. The attributes are sorted by most similar ($k = 0$) to least similar ($k = 15$). While there are some major similarities (short hair $k = 0$, eyeglasses $k = 5$), the lack of a dedicated constraint for finding hats shows the flaws with this approach. This example is a clear demonstration that without a dedicated cross-model constraint (i.e., using the DRE models), finding missing attributes is difficult.

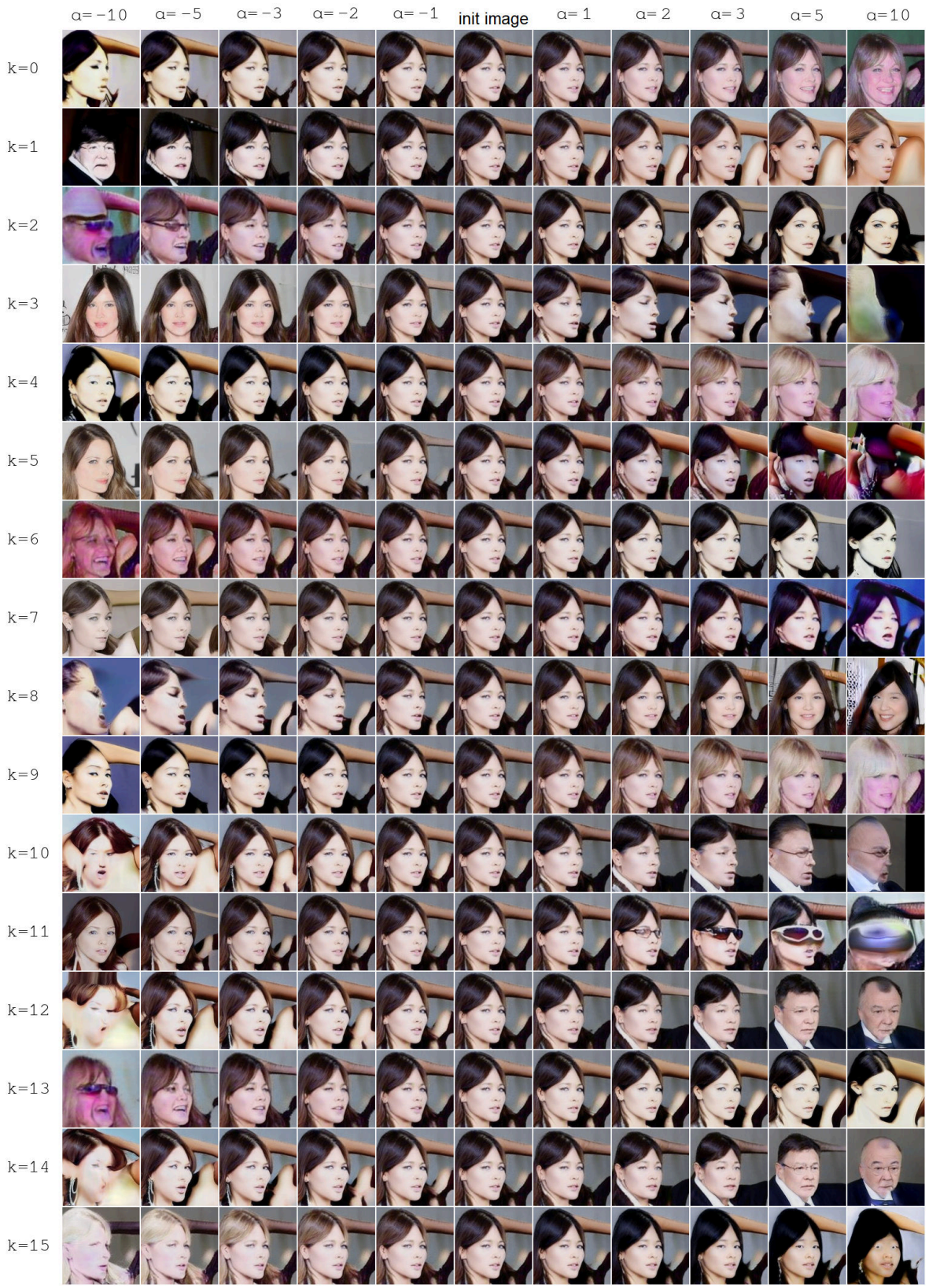

Figure 6. All 16 learned attributes of a Vanilla ResNet model.

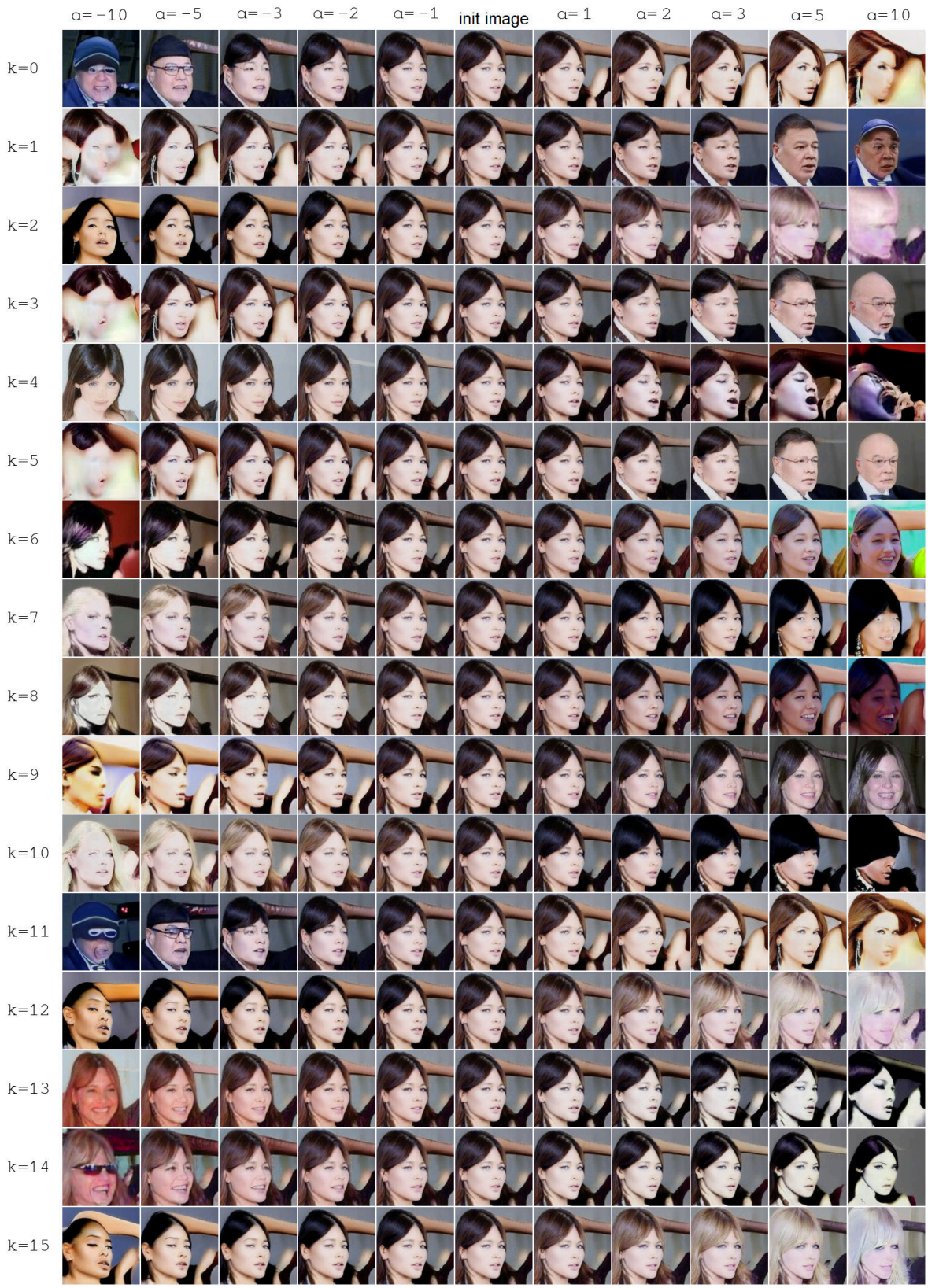

Figure 7. All 16 learned attributes of a robust ResNet model.

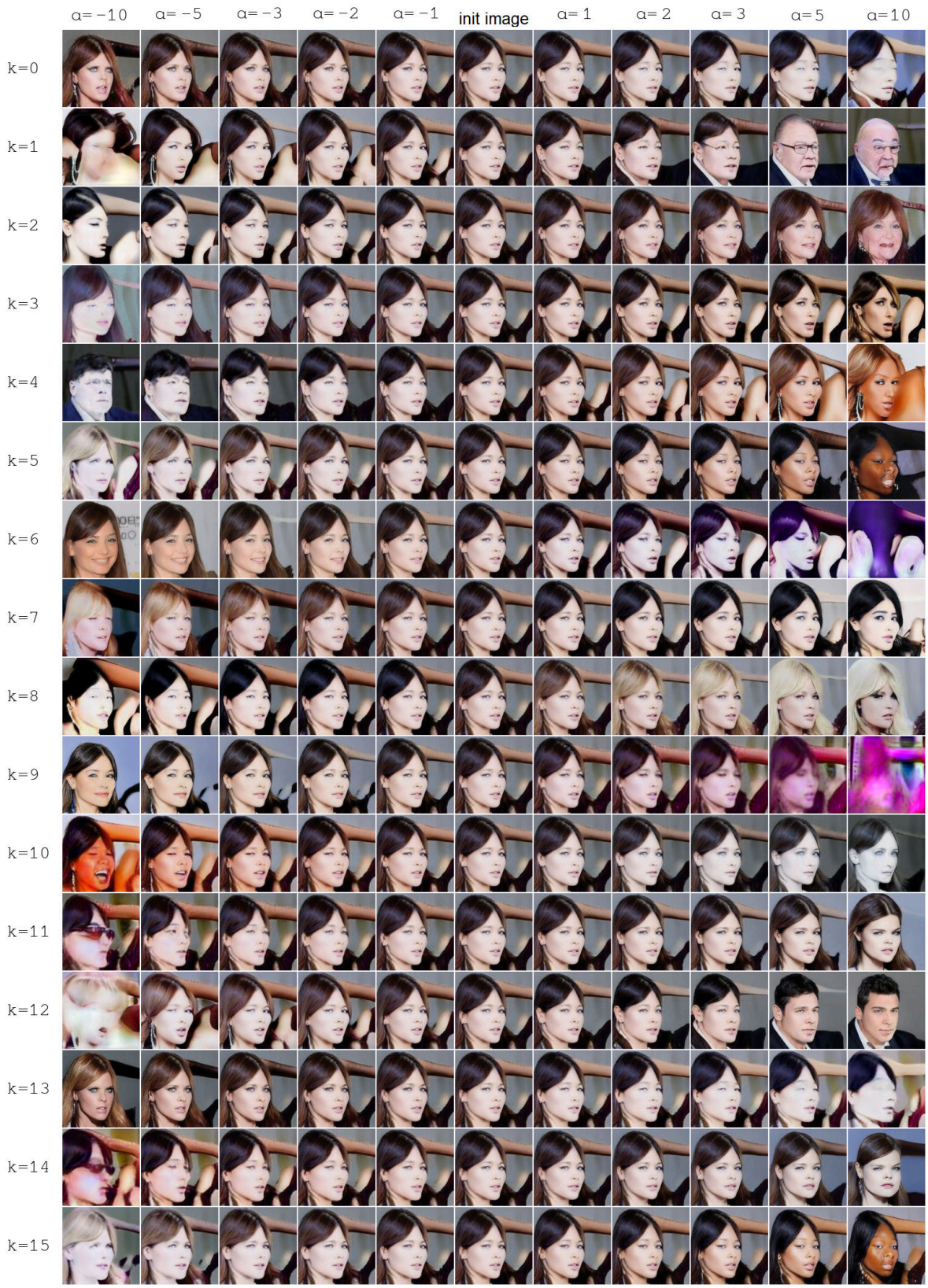

Figure 8. All 16 learned attributes of a clip ResNet model.

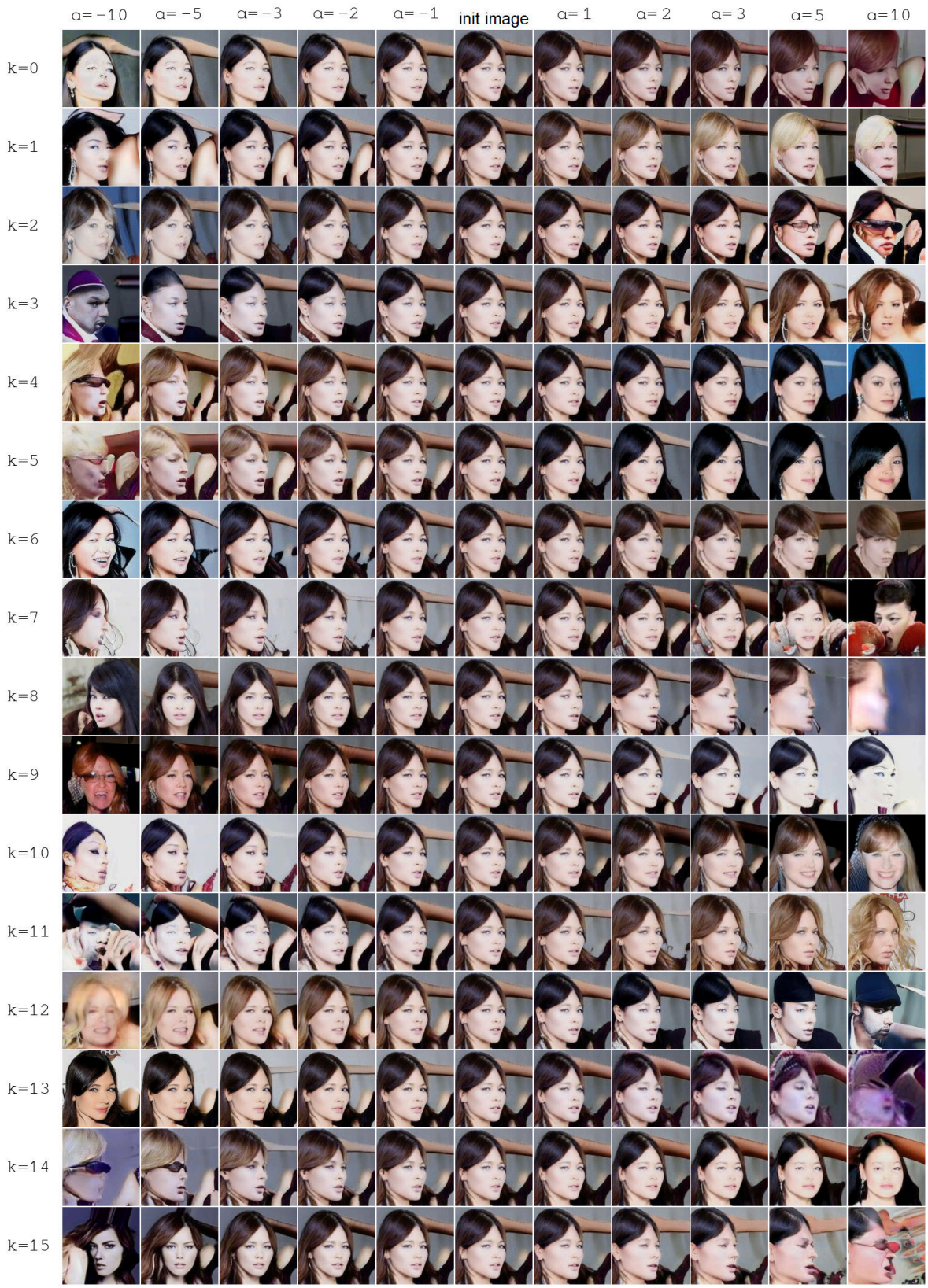

Figure 9. All 16 learned attributes of a Attribute Classifier ResNet model.

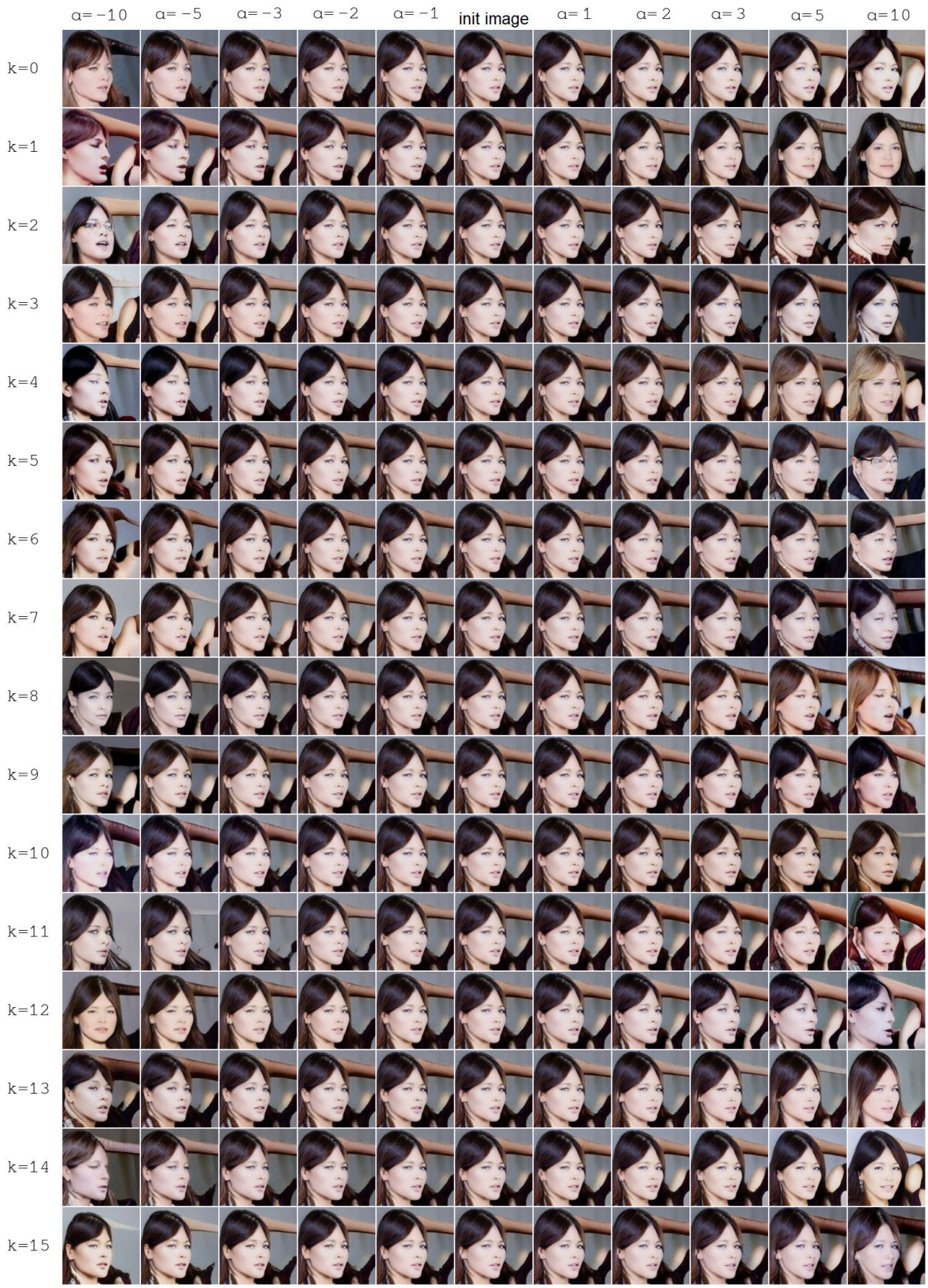

Figure 10. All 16 learned attributes of the original LatentCLR model (using global directions, rather than conditional).

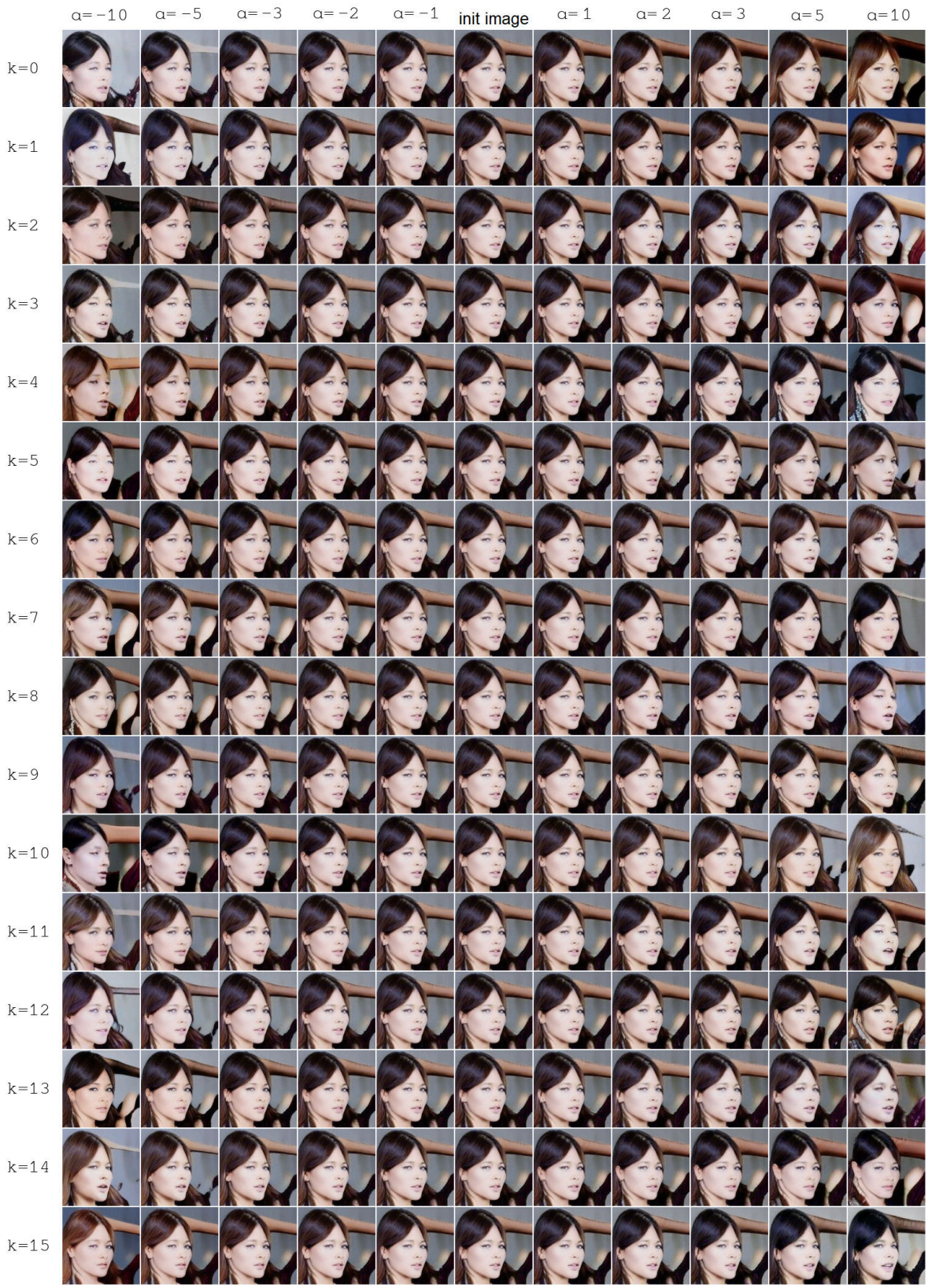

Figure 11. All 16 learned attributes of the Hessian method.

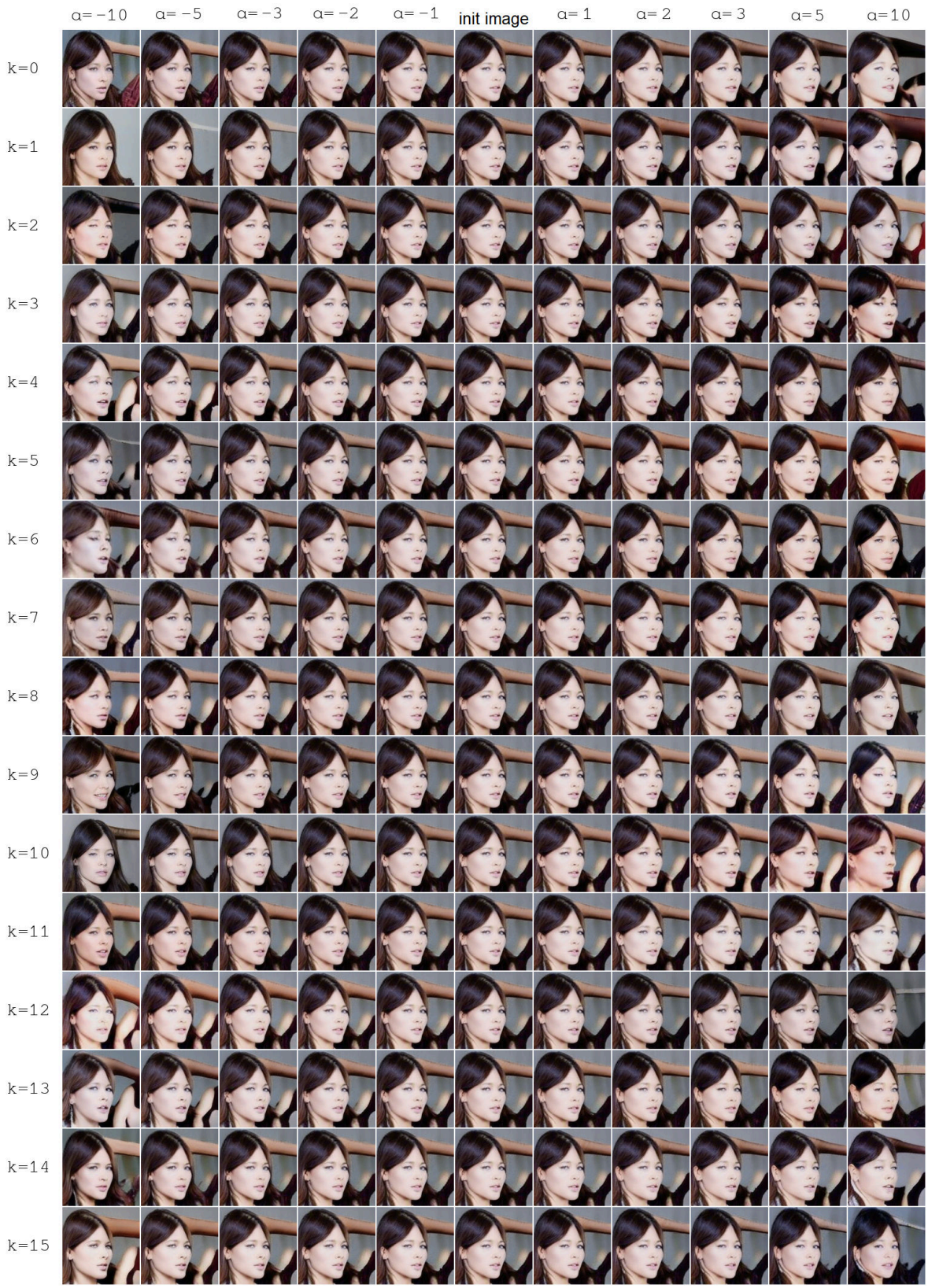

Figure 12. All 16 learned attributes of the Jacobian method.

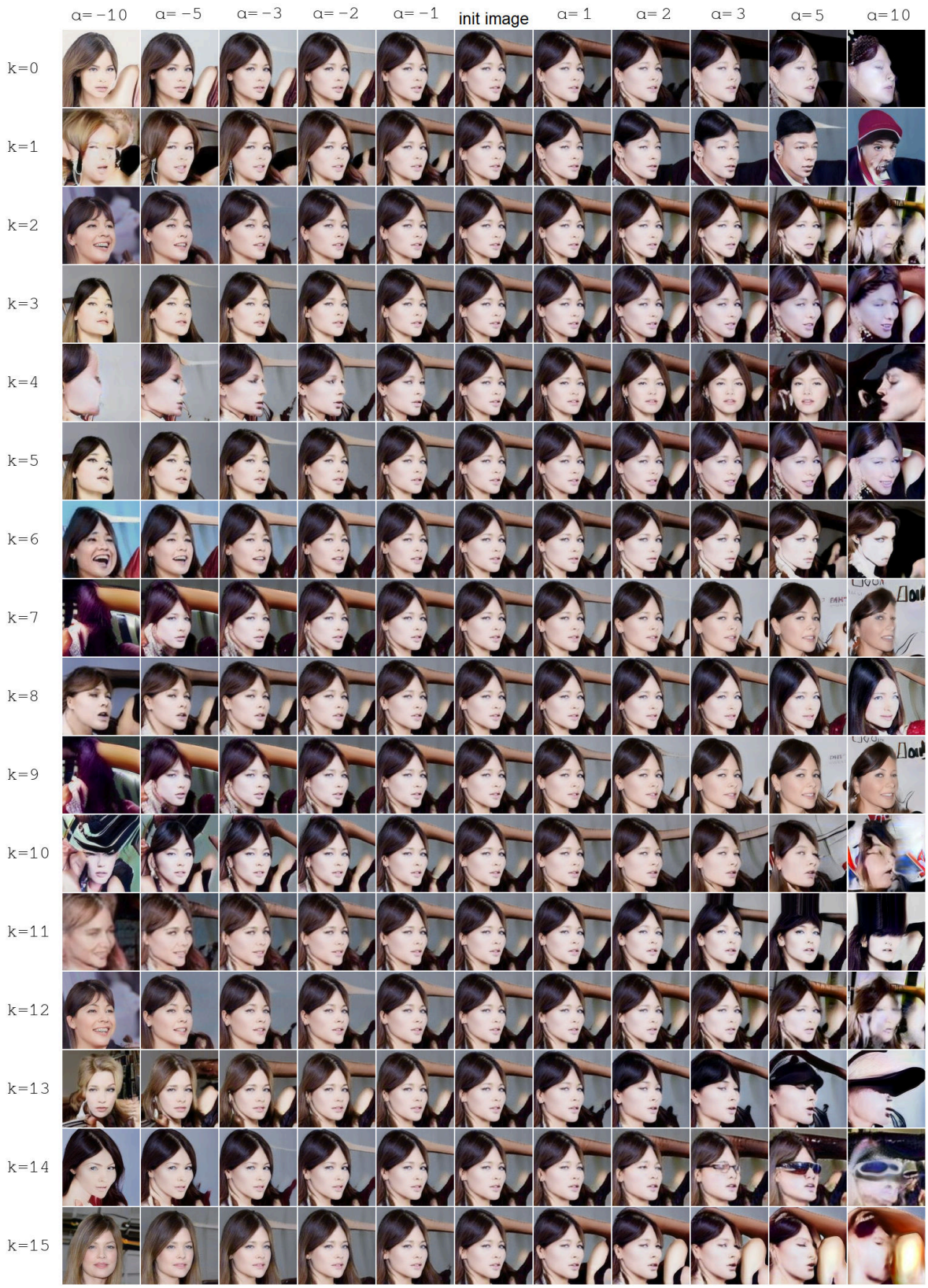

Figure 13. All 16 learned attributes of the original LatentCLR model with conditional directions.

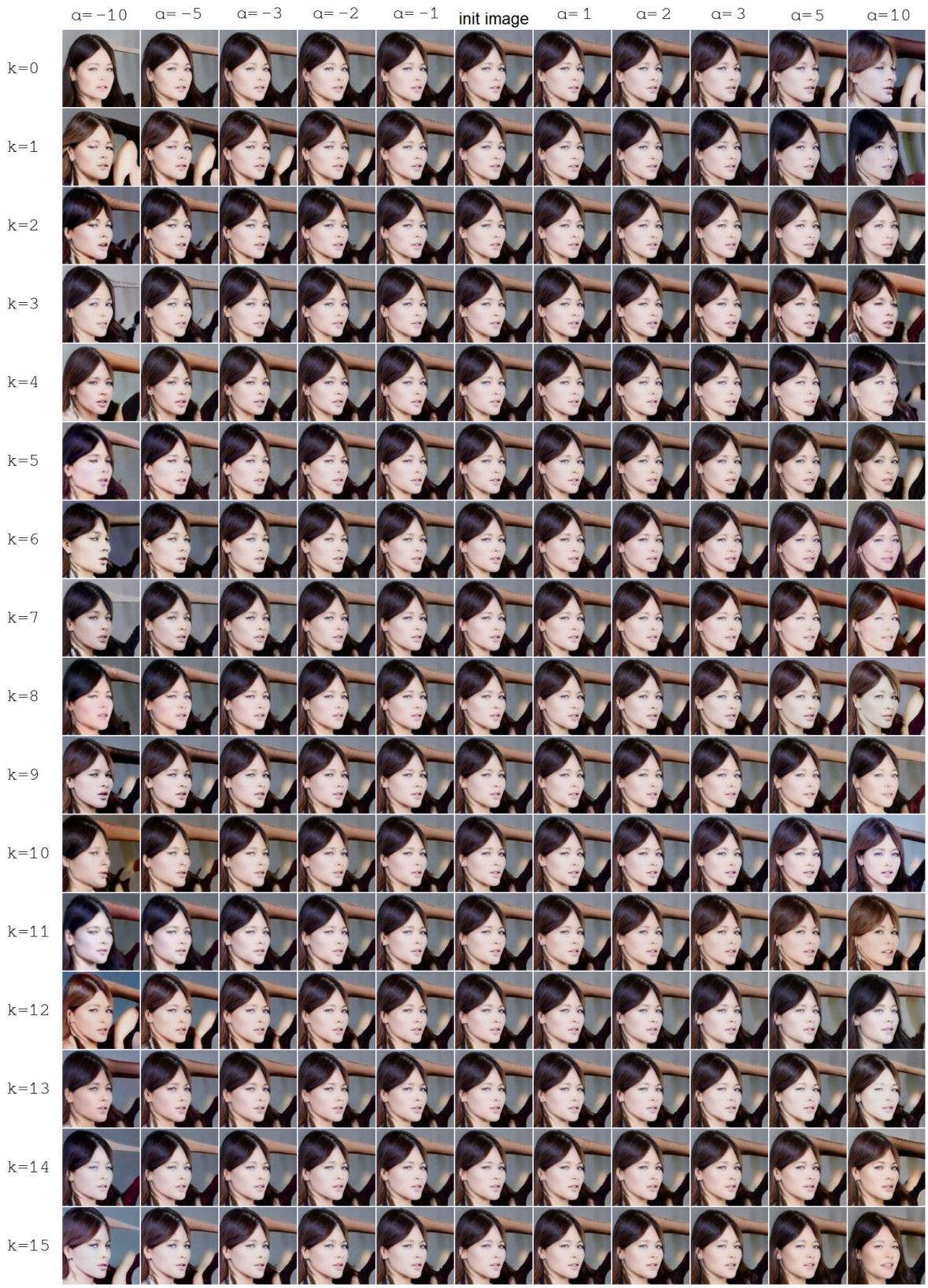

Figure 14. All 16 learned attributes of the Voynov method.

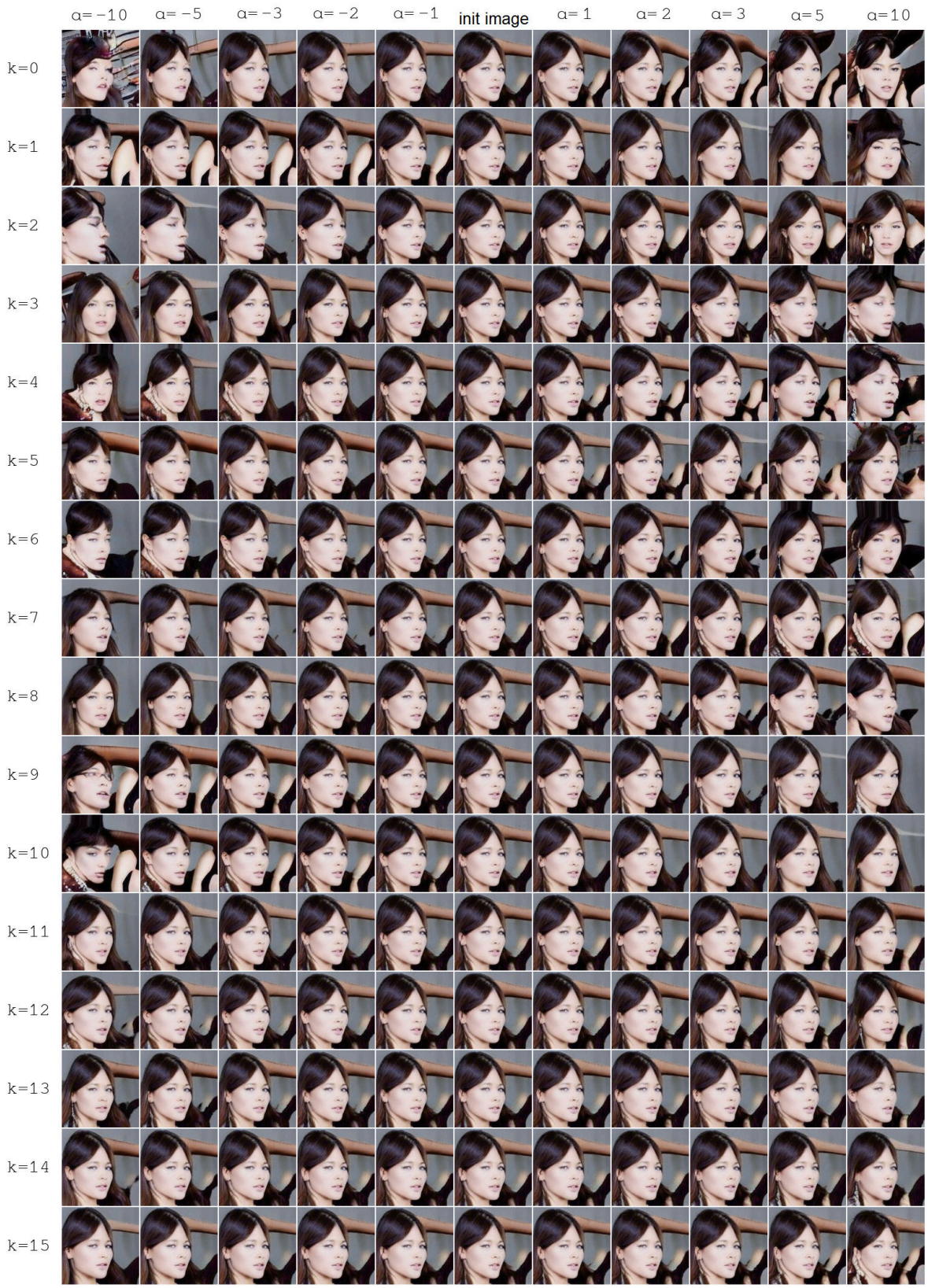

Figure 15. The top 16 learned attributes of the SeFa method.

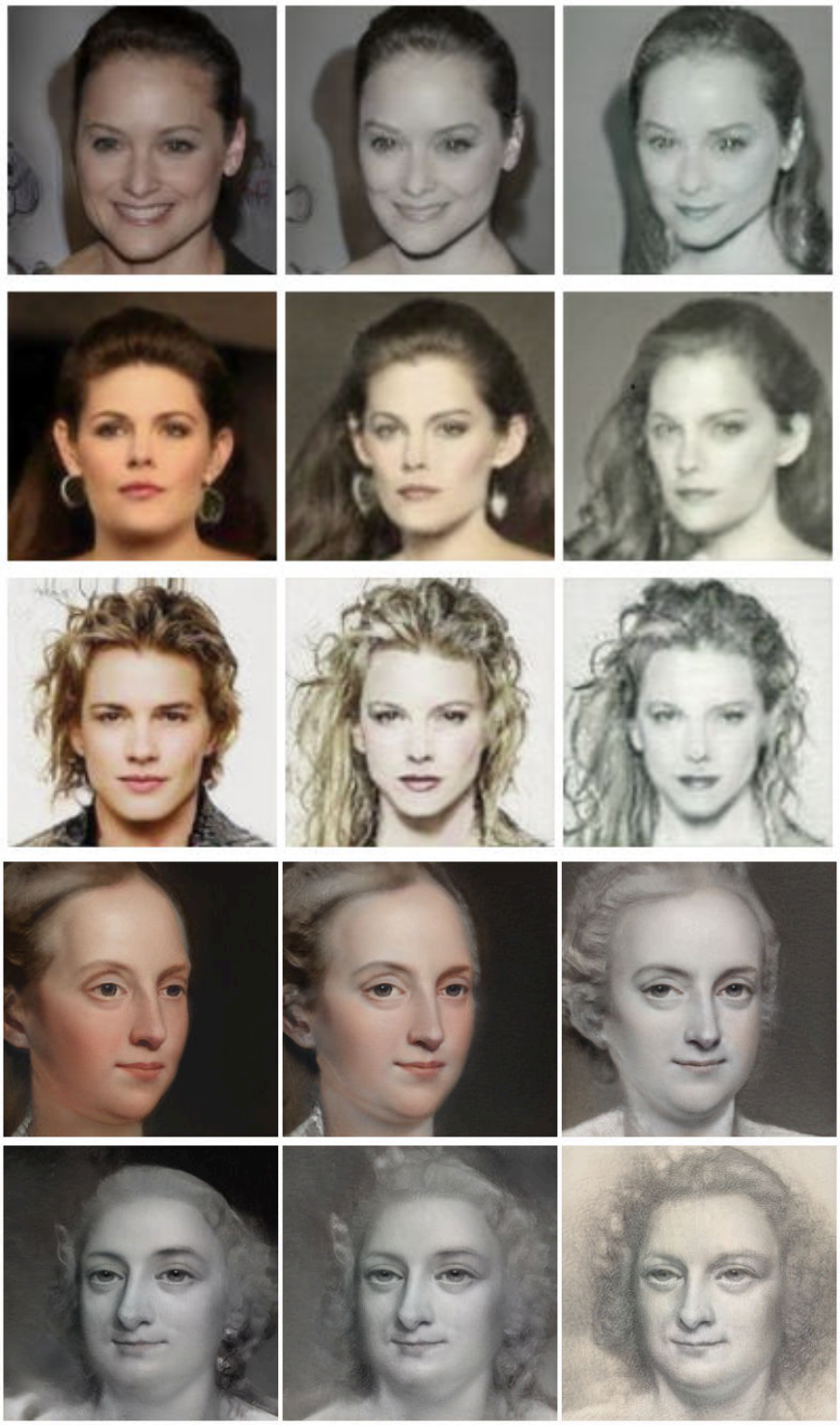

Figure 16. Examples of common sketch attribute between CelebA (Top) and Metfaces (Bottom).

Figure 17. Examples of common formal-wear attribute between CelebA (Top) and Metfaces (Bottom).

Figure 18. Examples of common white, curly hair attribute between CelebA (Top) and Metfaces (Bottom).

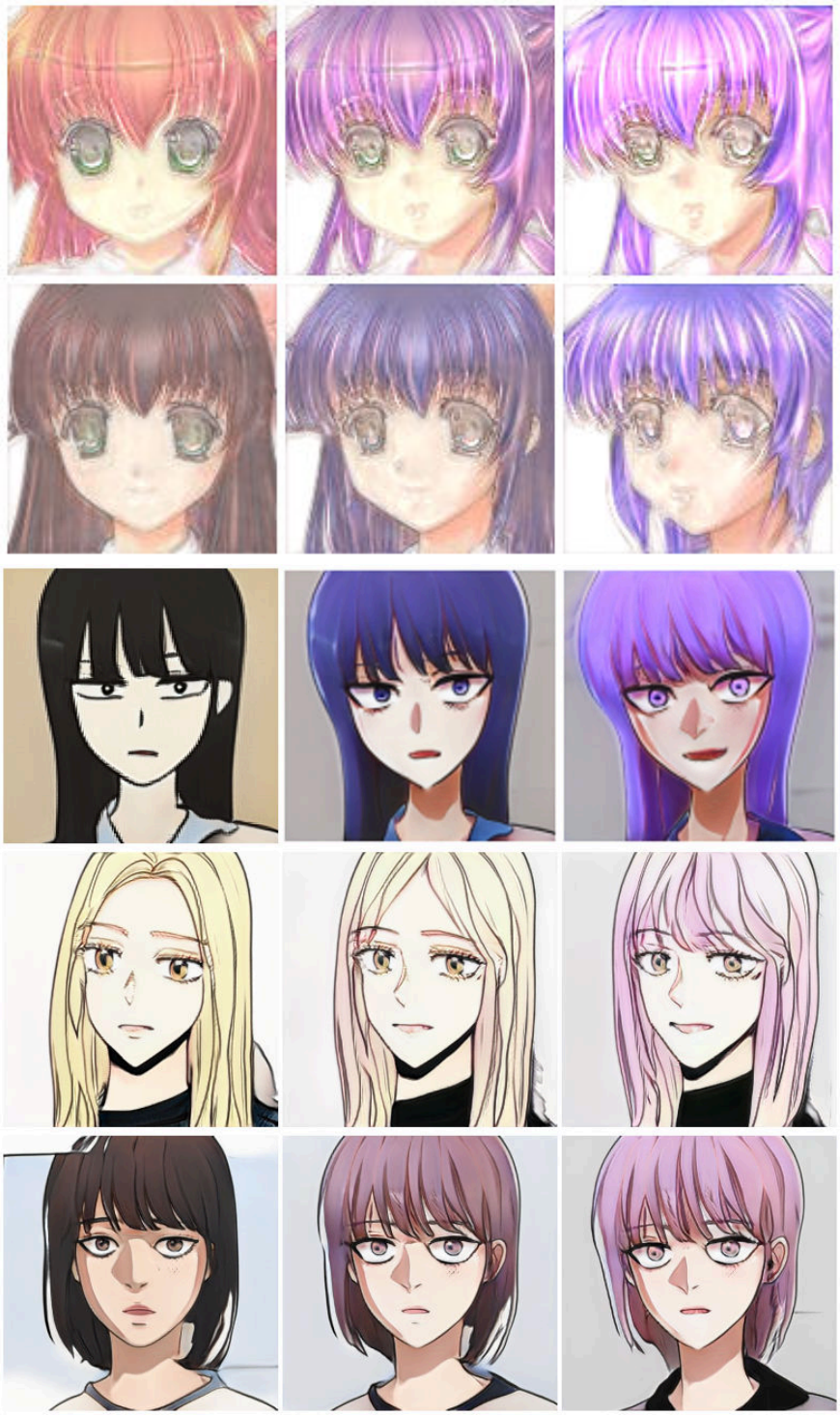

Figure 19. Examples of common purple hair attribute between Anime (Top) and Toon (Bottom).

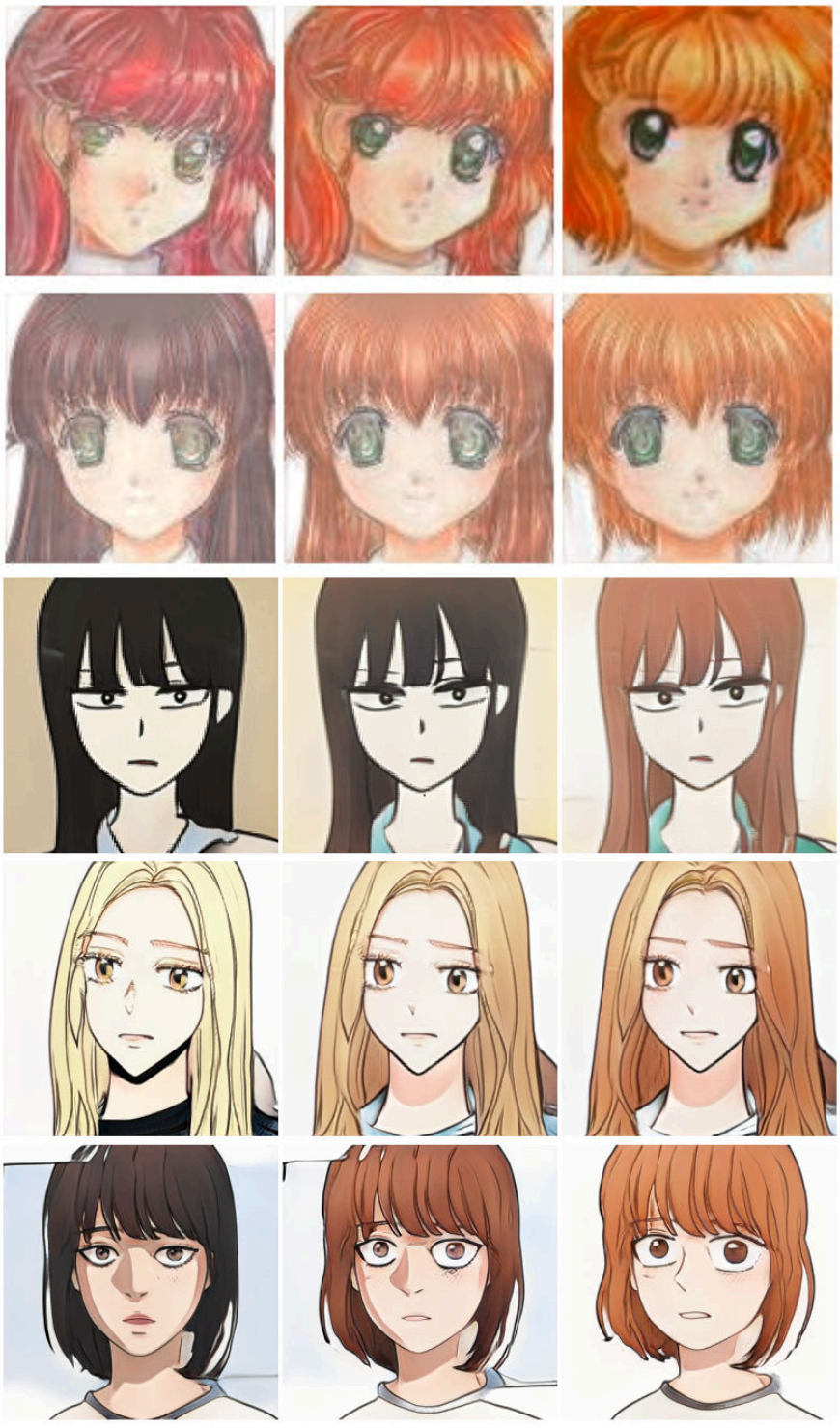

Figure 20. Examples of common orange/brown hair attribute between Anime (Top) and Toon (Bottom).

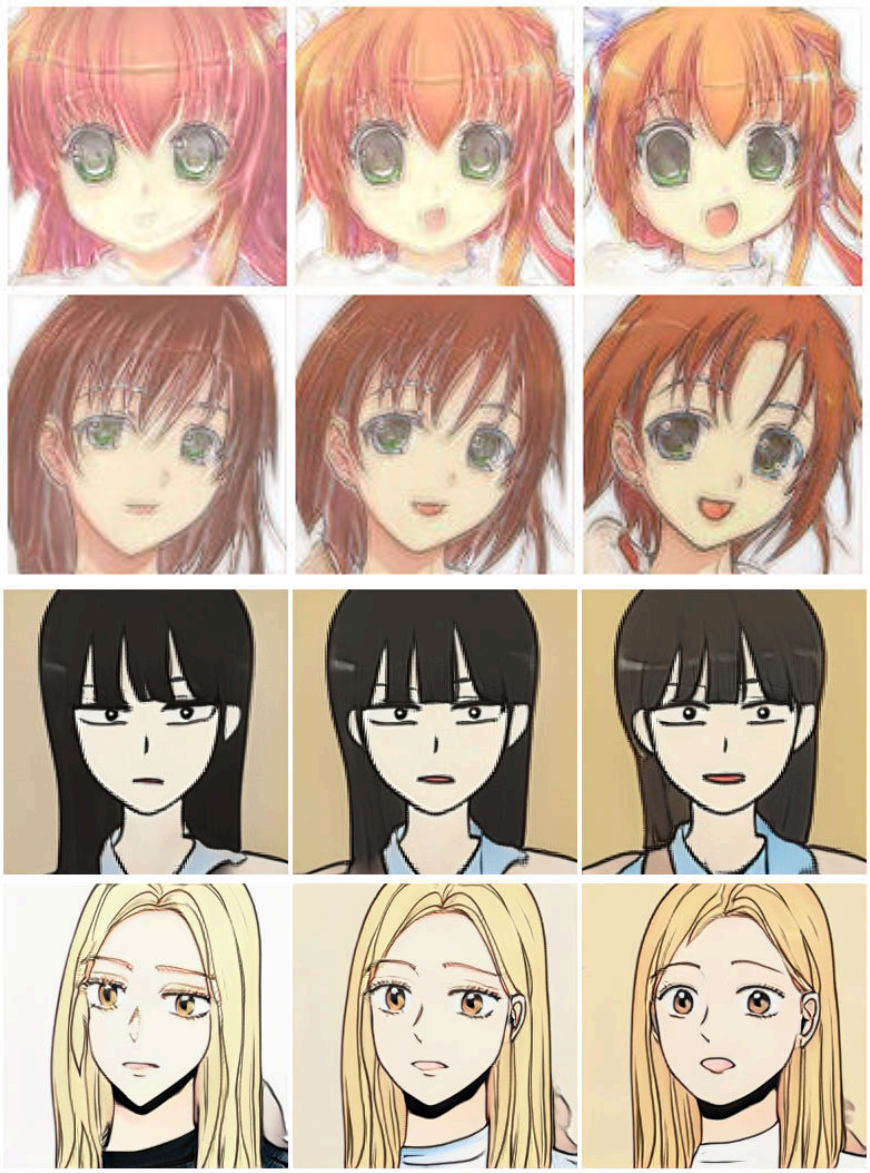

Figure 21. Examples of common open-mouth attribute between Anime (Top) and Toon (Bottom).

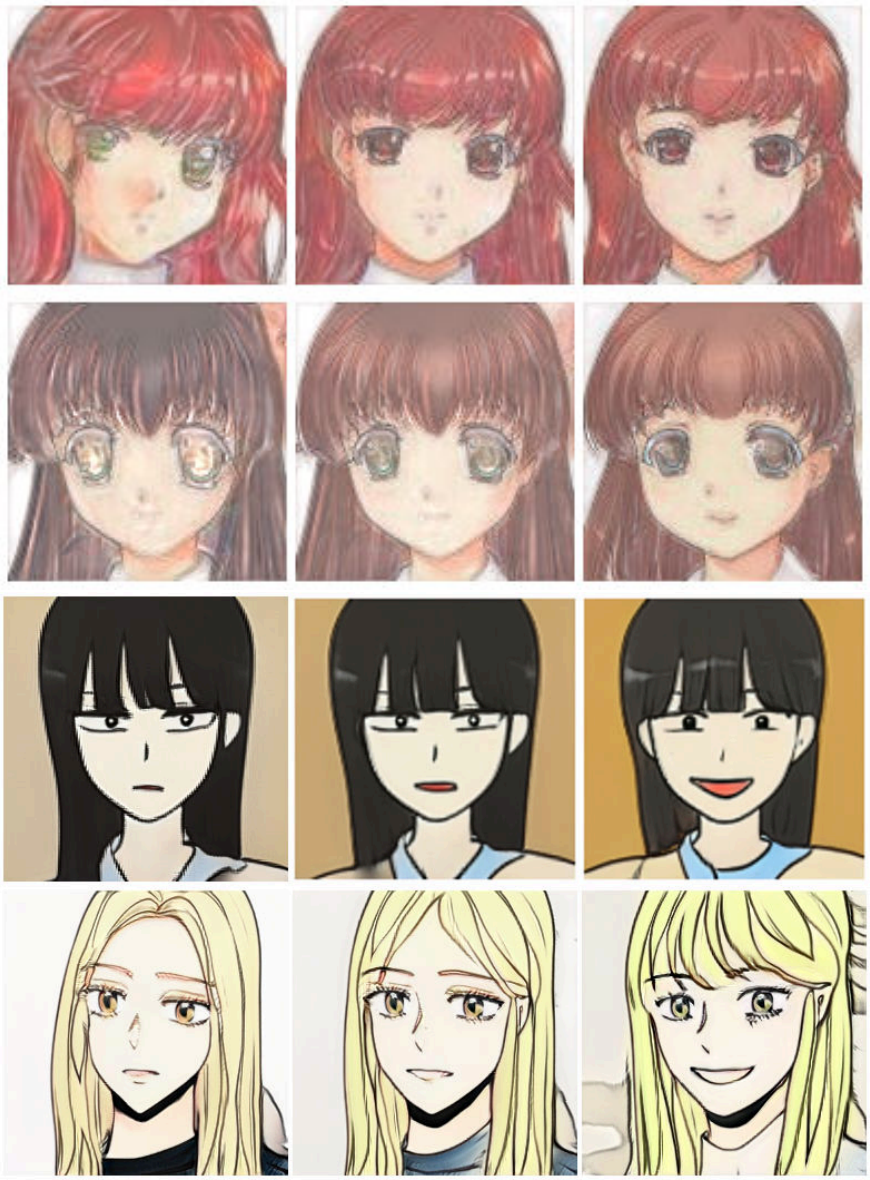

Figure 22. Examples of common smiling attribute between Anime (Top) and Toon (Bottom).

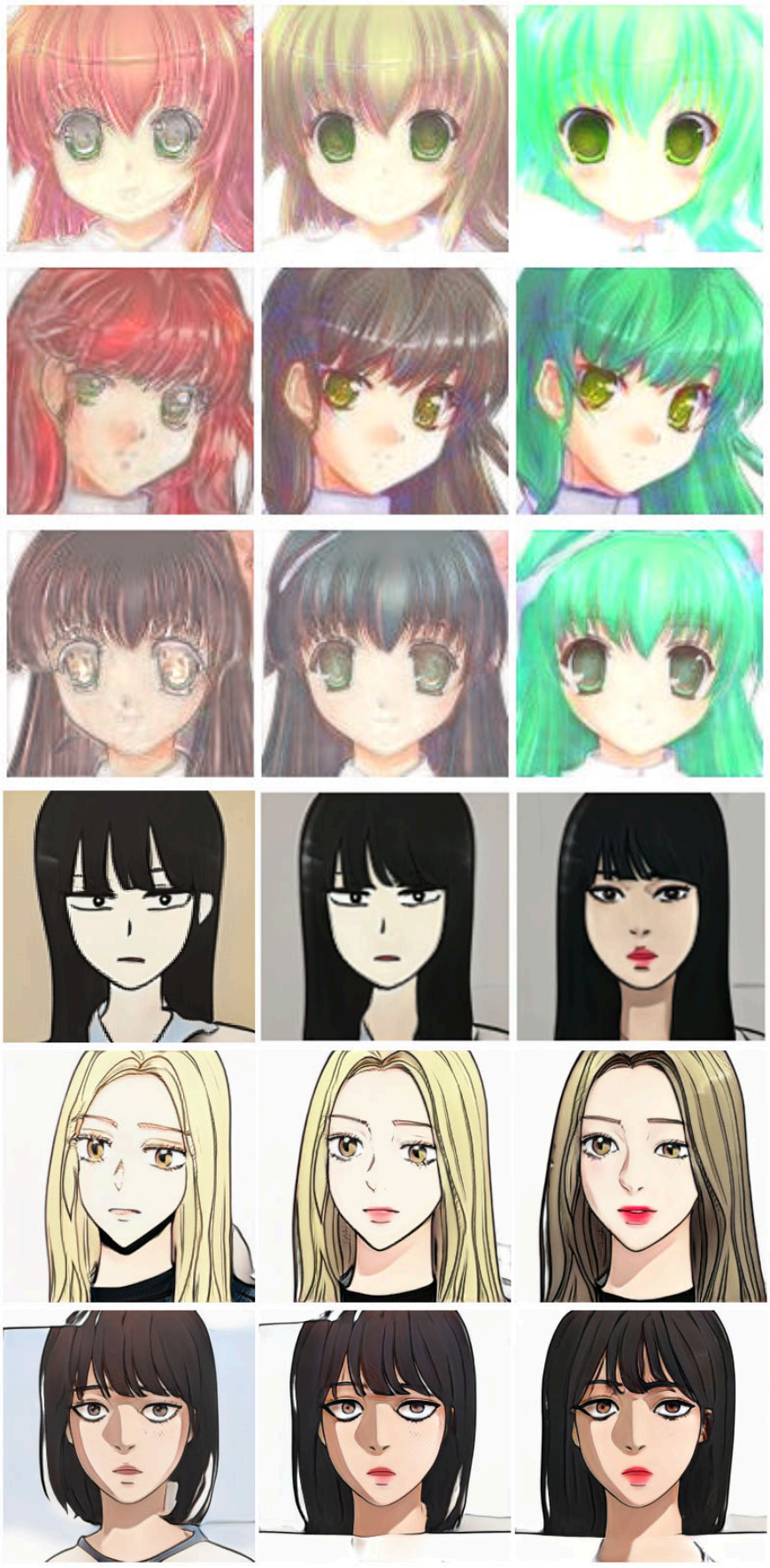

Figure 23. Examples of novel green hair attribute from Anime (Top) and the missing lipstick attribute from Toon (Bottom).

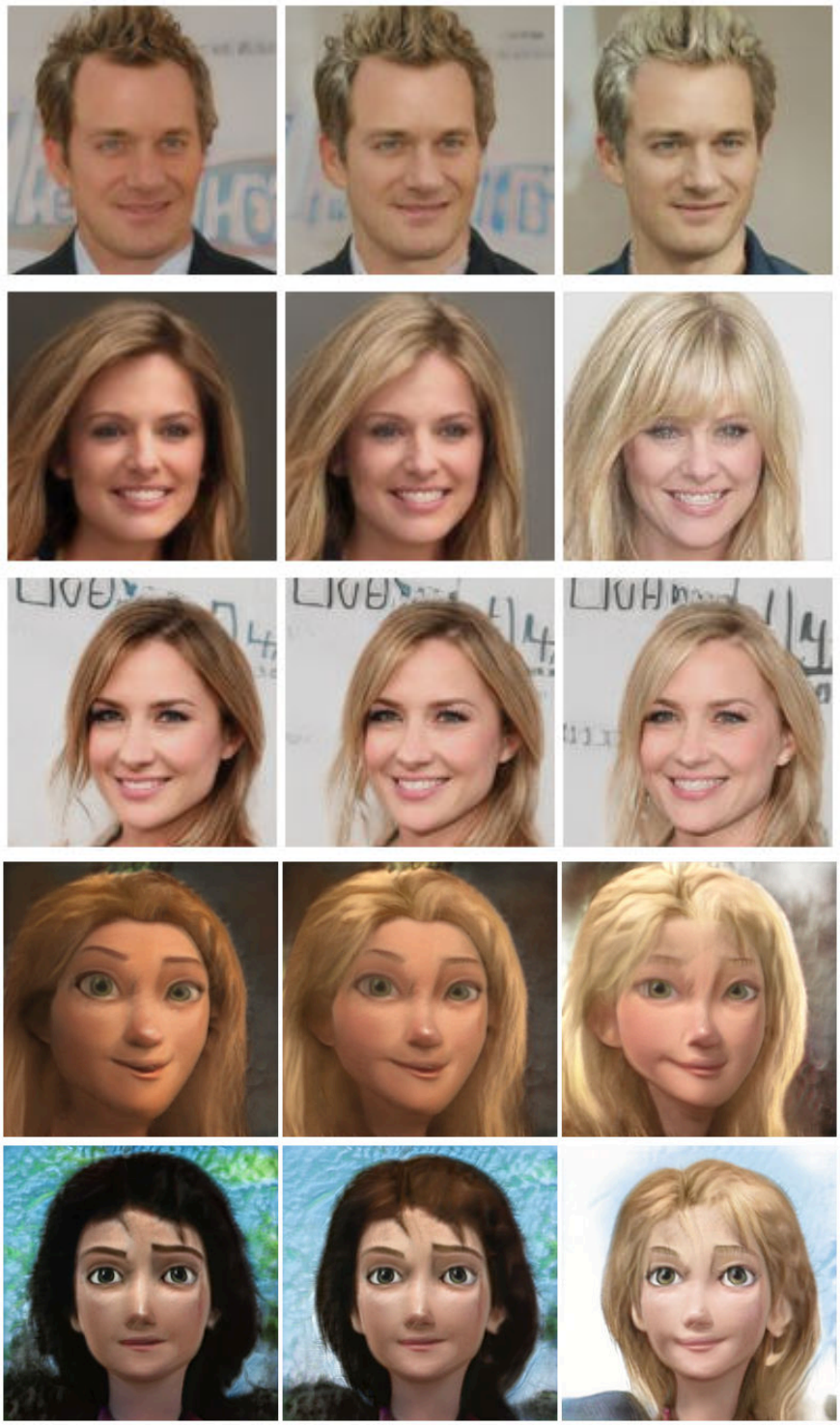

Figure 24. Examples of common blonde attribute between CelebA (Top) and Disney (Bottom).

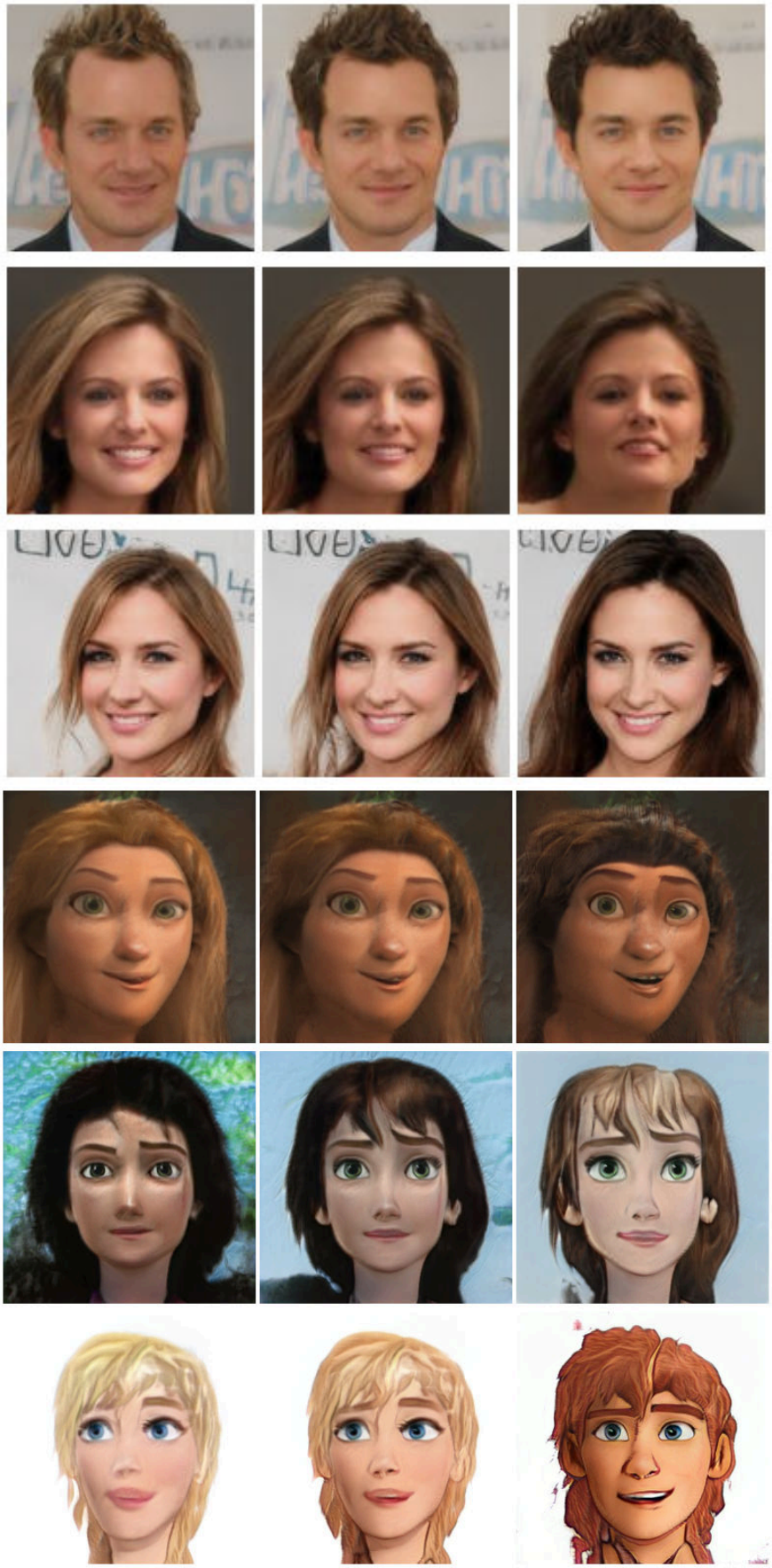

Figure 25. Examples of common brown hair attribute between CelebA (Top) and Disney (Bottom).

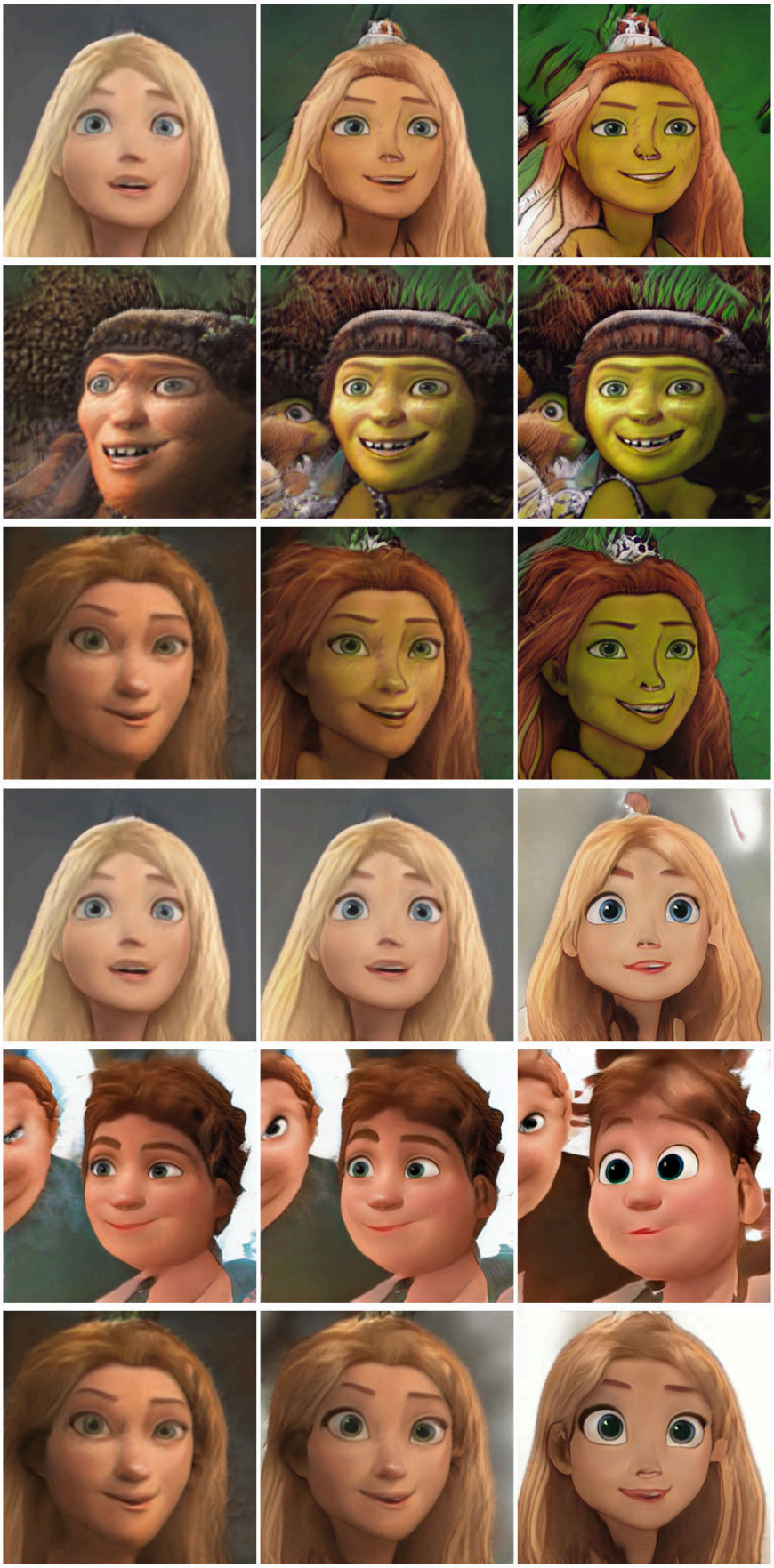

Figure 26. Two examples of novel Disney attributes: making princesses ogre-like and large cartoonish eyes.

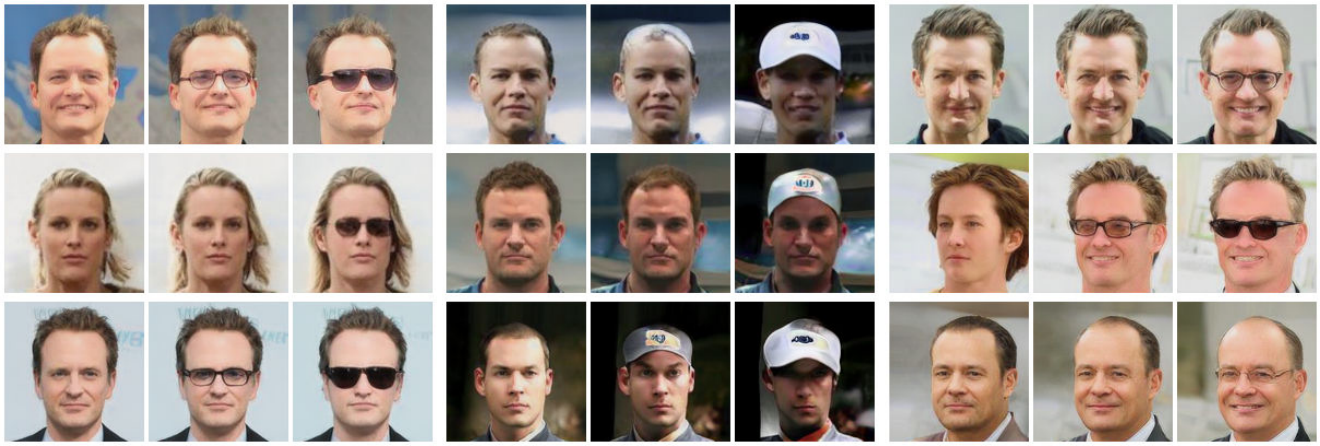

Figure 27. Examples of various missing attributes from full CelebA GAN against three different attribute splits: **(Left 3)** the missing eyeglass attribute, **(Middle 3)** the missing hats attribute and **(Right 3)** the missing eyeglass, smiling, and attempting to identify the necktie attribute.

	Full CelebA	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA		0.143	0.143	0.045	0.111	0.063	0.278	0.189
Female	0.000		0.167	1.000	0.167	0.200	0.170	0.417
Male	0.000	0.059		0.250	0.111	0.250	0.188	0.303
No Hats	0.000	0.056	0.111		0.083	0.067	0.188	0.267
No Glasses	0.000	0.059	0.125	0.333		0.042	0.306	0.203
No Beards	0.000	0.333	0.083	0.043	0.063		0.185	0.107
No Beard No Hats	0.000	0.143	0.100	0.125	0.143	0.125		0.511
No Glasses No Smiles No Ties	0.000	0.053	0.333	0.250	0.500	0.059	0.096	

Table 5. The full results for the recovery scores ($\mathcal{R}_{\text{score}}$) of the SeFa method.

	Full CelebA	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	-	0.478	0.536	0.086	0.390	0.388	0.120	0.287
Female	0	-	0.246	0.048	0.050	0.046	0.048	0.279
Male	0	0.586	-	0.124	0.333	0.733	0.384	0.405
No Hats	0	0.251	0.240	-	0.097	0.180	0.108	0.313
No Glasses	0	0.585	0.542	0.048	-	0.114	0.100	0.218
No Beards	0	0.290	0.583	0.069	0.080	-	0.066	0.271
No Beard No Hats	0	0.373	0.396	0.034	0.189	0.049	-	0.297
No Glasses No Smiles No Ties	0	0.448	0.667	0.055	0.033	0.537	0.269	-

Table 6. The full results for the recovery scores ($\mathcal{R}_{\text{score}}$) of the Jacobian loss.

	Full CelebA	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	-	1.000	1.000	0.056	0.167	0.167	0.096	0.407
Female	0	-	0.200	0.056	0.045	0.036	0.057	0.364
Male	0	0.333	-	0.077	1.000	1.000	0.300	0.300
No Hats	0	0.500	0.250	-	0.083	0.071	0.042	0.370
No Glasses	0	0.333	0.250	0.083	-	0.111	0.071	0.150
No Beards	0	0.125	0.167	0.071	0.063	-	0.046	0.218
No Beard No Hats	0	0.167	0.333	0.032	0.071	0.053	-	0.375
No Glasses No Smiles No Ties	0	1.000	0.333	0.200	0.048	0.143	0.102	-

Table 7. The full results for the recovery scores ($\mathcal{R}_{\text{score}}$) of the Hessian loss.

	Full CelebA	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	-	1.000	1.000	0.250	0.333	0.200	0.153	0.537
Female	0	-	0.250	0.048	0.034	0.050	0.044	0.137
Male	0	1.000	-	0.143	0.500	0.333	0.267	0.242
No Hats	0	0.250	1.000	-	0.125	0.167	0.077	0.158
No Glasses	0	0.333	1.000	0.038	-	0.250	0.525	0.153
No Beards	0	0.125	0.500	0.045	0.063	-	0.049	0.381
No Beard No Hats	0	1.000	1.000	0.043	0.167	0.091	-	0.389
No Glasses No Smiles No Ties	0	1.000	0.500	0.067	0.033	0.125	0.072	-

Table 8. The full results for the recovery scores ($\mathcal{R}_{\text{score}}$) of the LatentCLR loss.

	Full CelebA	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	-	0.500	1.000	0.333	0.050	0.500	0.153	0.259
Female	0	-	0.143	0.045	0.167	0.038	0.061	0.410
Male	0	0.500	-	0.250	0.333	0.333	0.306	0.256
No Hats	0	0.500	0.143	-	0.091	0.500	0.139	0.511
No Glasses	0	0.333	1.000	0.050	-	0.167	0.094	0.377
No Beards	0	0.167	0.250	0.167	0.091	-	0.052	0.370
No Beard No Hats	0	0.500	0.500	0.034	0.050	0.034	-	0.386
No Glasses No Smiles No Ties	0	0.143	0.500	0.143	0.029	1.000	0.286	-

Table 9. The full results for the recovery scores ($\mathcal{R}_{\text{score}}$) of the voynov loss.

	Full CelebA	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	-	1.000	1.000	0.333	0.500	0.200	0.167	0.465
Female	0	-	0.333	0.053	0.067	0.034	0.046	0.381
Male	0	1	-	0.059	0.250	0.083	0.082	0.521
No Hats	0	1	0.5	-	0.143	0.071	0.062	0.460
No Glasses	0	1	0.5	0.059	-	0.091	0.081	0.377
No Beards	0	1	1	1	0.091	-	0.154	0.378
No Beard No Hats	0	1	1	0.033	0.05	0.042	-	0.382
No Glasses No Smiles No Ties	0	1	1	0.083	0.029	0.333	0.122	-

Table 10. The full results for the recovery scores ($\mathcal{R}_{\text{score}}$) of the ImageNet ResNet.

	Full CelebA	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	-	1.000	0.333	1.000	0.091	1.000	0.525	0.390
Female	0	-	0.200	0.042	0.059	0.038	0.036	0.198
Male	0	1	-	0.063	0.143	0.056	0.306	0.492
No Hats	0	1	0.333	-	0.333	1.000	0.516	0.365
No Glasses	0	1	0.333	0.1	-	0.500	0.270	0.357
No Beards	0	1	1	0.053	0.111	-	0.042	0.211
No Beard No Hats	0	1	1	0.033	0.071	0.048	-	0.212
No Glasses No Smiles No Ties	0	1	1	1	0.03	0.5	0.274	-

Table 11. The full results for the recovery scores ($\mathcal{R}_{\text{score}}$) of the Attribute Classifier ResNet.

	Full CelebA	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	-	1.000	1.000	0.250	1.000	0.063	0.183	0.401
Female	0	-	0.333	0.500	0.042	0.048	0.047	0.363
Male	0	0.5	-	0.077	0.500	0.111	0.563	0.419
No Hats	0	1	1	-	1.000	0.143	0.517	0.423
No Glasses	0	1	0.333	0.034	-	0.333	0.200	0.361
No Beards	0	1	1	0.143	0.045	-	0.113	0.370
No Beard No Hats No Glasses	0	1	1	0.028	0.5	0.038	-	0.376
No Smiles No Ties	0	1	1	0.056	0.033	1	0.556	-

Table 12. The full results for the recovery scores ($\mathcal{R}_{\text{score}}$) of the Robust ResNet.

	Full CelebA	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	-	1.000	1.000	1.000	0.200	0.200	0.108	0.383
Female	0	-	0.250	0.028	0.200	0.063	0.131	0.194
Male	0	0.143	-	0.028	0.200	0.063	0.131	0.055
No Hats	0	1	1	-	0.500	0.125	0.133	0.231
No Glasses	0	1	1	0.037	-	0.100	0.148	0.397
No Beards	0	1	1	0.083	0.333	-	0.098	0.373
No Beard No Hats No Glasses	0	1	1	0.1	0.2	0.029	-	0.371
No Smiles No Ties	0	1	1	0.063	0.125	0.167	0.205	-

Table 13. The full results for the recovery scores ($\mathcal{R}_{\text{score}}$) of the CLIP ResNet.

	Full CelebA	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	-	1.000	1.000	0.200	0.091	0.050	0.276	0.363
Female	0	-	0.333	0.040	0.045	0.056	0.054	0.140
Male	0	1	-	0.500	0.111	0.063	0.140	0.199
No Hats	0	0.5	1	-	0.200	0.040	0.076	0.369
No Glasses	0	0.333	1	0.111	-	0.063	0.042	0.388
No Beards	0	0.143	0.143	0.125	0.1	-	0.086	0.209
No Beard No Hats No Glasses	0	0.143	0.5	0.031	0.053	0.032	-	0.192
No Smiles No Ties	0	1	1	0.033	0.125	0.125	0.108	-

Table 14. The full results for the recovery scores ($\mathcal{R}_{\text{score}}$) of the ImageNet ViT.

	Full CelebA	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	-	0.500	1.000	0.027	1.000	0.067	0.264	0.387
Female	0	-	0.200	0.034	0.032	0.036	0.048	0.188
Male	0	0.333	-	0.091	0.125	0.045	0.229	0.232
No Hats	0	1	1	-	0.063	0.067	0.170	0.365
No Glasses	0	1	0.5	0.067	-	0.053	0.066	0.189
No Beards	0	0.5	1	0.091	0.333	-	0.073	0.127
No Beard No Hats	0	0.2	0.5	0.029	0.056	0.038	-	0.369
No Glasses No Smiles No Ties	0	1	1	0.032	0.125	0.037	0.107	-

Table 15. The full results for the recovery scores ($\mathcal{R}_{\text{score}}$) of the CLIP ViT.

	Full CelebA	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	-	0.200	0.250	0.100	0.077	0.077	0.156	0.068
Female	0	-	0.200	0.038	0.056	0.033	0.044	0.119
Male	0	1	-	0.053	0.143	0.042	0.104	0.511
No Hats	0	0.091	1	-	0.063	0.028	0.046	0.194
No Glasses	0	0.25	1	0.067	-	0.034	0.047	0.081
No Beards	0	0.5	1	0.045	0.083	-	0.072	0.356
No Beard No Hats	0	0.25	0.333	0.067	0.059	0.034	-	0.068
No Glasses No Smiles No Ties	0	1	0.5	0.027	0.091	0.043	0.163	-

Table 16. The full results for the recovery scores ($\mathcal{R}_{\text{score}}$) of the MAE ViT.

	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	0.413	0.458	0.3715	0.374	0.314	0.387	0.355
Female	-	0.329	0.3615	0.320	0.352	0.360	0.334
Male	-	-	0.3732	0.426	0.364	0.352	0.381
No Hats	-	-	-	0.400	0.300	0.325	0.379
No Glasses	-	-	-	-	0.336	0.305	0.343
No Beards	-	-	-	-	-	0.302	0.295
No Beard No Hats	-	-	-	-	-	-	0.341

Table 17. The full results for the alignment scores ($\mathcal{A}_{\text{score}}$) of the SeFa method.

	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	0.475	0.489	0.652	0.598	0.618	0.615	0.525
Female	-	0.367	0.523	0.457	0.559	0.559	0.345
Male	-	-	0.466	0.401	0.372	0.431	0.450
No Hats	-	-	-	0.512	0.620	0.505	0.518
No Glasses	-	-	-	-	0.556	0.502	0.503
No Beards	-	-	-	-	-	0.563	0.477
No Beard No Hats	-	-	-	-	-	-	0.439

Table 18. The full results for the alignment scores ($\mathcal{A}_{\text{score}}$) of the Hessian loss.

	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	0.519	0.511	0.556	0.533	0.512	0.593	0.579
Female	-	0.412	0.452	0.513	0.550	0.613	0.388
Male	-	-	0.373	0.474	0.425	0.448	0.426
No Hats	-	-	-	0.460	0.485	0.576	0.482
No Glasses	-	-	-	-	0.484	0.630	0.491
No Beards	-	-	-	-	-	0.550	0.501
No Beard No Hats	-	-	-	-	-	-	0.496

Table 19. The full results for the alignment scores ($\mathcal{A}_{\text{score}}$) of the LatentCLR loss.

	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	0.566	0.477	0.567	0.555	0.570	0.562	0.513
Female	-	0.430	0.606	0.609	0.577	0.592	0.505
Male	-	-	0.508	0.503	0.429	0.456	0.453
No Hats	-	-		0.596	0.553	0.583	0.553
No Glasses	-	-	-	-	0.572	0.559	0.564
No Beards	-	-	-	-	-	0.616	0.567
No Beard No Hats	-	-	-	-	-	-	0.499

Table 20. The full results for the alignment scores ($\mathcal{A}_{\text{score}}$) of the Voynov method.

	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	0.523	0.452	0.505	0.528	0.519	0.491	0.495
Female	-	0.413	0.481	0.522	0.490	0.511	0.449
Male	-	-	0.488	0.451	0.443	0.409	0.384
No Hats	-	-	-	0.498	0.515	0.523	0.465
No Glasses	-	-	-	-	0.497	0.517	0.503
No Beards	-	-	-	-	-	0.544	0.484
No Beard No Hats	-	-	-	-	-	-	0.512

Table 21. The full results for the alignment scores ($\mathcal{A}_{\text{score}}$) of the Jacobian loss.

	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	0.457	0.403	0.740	0.792	0.461	0.643	0.489
Female	-	0.409	0.651	0.720	0.599	0.483	0.444
Male	-	-	0.508	0.377	0.328	0.360	0.390
No Hats	-	-	-	0.611	0.778	0.414	0.560
No Glasses	-	-	-	-	0.698	0.659	0.649
No Beards	-	-	-	-	-	0.652	0.556
No Beard No Hats	-	-	-	-	-	-	0.492

Table 22. The full results for the alignment scores ($\mathcal{A}_{\text{score}}$) of the ImageNet Trained ResNet.

	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	0.069	0.272	0.419	0.494	0.556	0.543	0.014
Female	-	0.155	0.411	0.230	0.191	0.354	0.056
Male	-	-	0.248	0.392	0.086	0.338	0.231
No Hats	-	-	-	0.370	0.279	0.346	0.569
No Glasses	-	-	-	-	0.494	0.405	0.447
No Beards	-	-	-	-	-	0.331	0.510
No Beard No Hats	-	-	-	-	-	-	0.403

Table 23. The full results for the alignment scores ($\mathcal{A}_{\text{score}}$) of the Attribute Classifier ResNet.

	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	0.615	0.357	0.750	0.825	0.619	0.803	0.649
Female	-	0.439	0.643	0.644	0.404	0.565	0.525
Male	-	-	0.384	0.444	0.417	0.175	0.289
No Hats	-	-	-	0.508	0.310	0.744	0.641
No Glasses	-	-	-	-	0.717	0.682	0.557
No Beards	-	-	-	-	-	0.568	0.495
No Beard No Hats	-	-	-	-	-	-	0.474

Table 24. The full results for the alignment scores ($\mathcal{A}_{\text{score}}$) of the Robust ResNet.

	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	0.753	0.451	0.772	0.791	0.894	0.580	0.656
Female	-	0.283	0.733	0.684	0.639	0.474	0.337
Male	-	-	0.371	0.435	0.396	0.337	0.314
No Hats	-	-	-	0.815	0.679	0.715	0.505
No Glasses	-	-	-	-	0.748	0.608	0.623
No Beards	-	-	-	-	-	0.706	0.507
No Beard No Hats	-	-	-	-	-	-	0.723

Table 25. The full results for the alignment scores ($\mathcal{A}_{\text{score}}$) of the CLIP ResNet.

	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	0.308	0.417	0.433	0.486	0.455	0.409	0.416
Female	-	0.251	0.348	0.476	0.468	0.495	0.365
Male	-	-	0.168	0.275	0.304	0.331	0.363
No Hats	-	-	-	0.409	0.484	0.499	0.390
No Glasses	-	-	-	-	0.347	0.363	0.417
No Beards	-	-	-	-	-	0.344	0.164
No Beard No Hats	-	-	-	-	-	-	0.363

Table 26. The full results for the alignment scores ($\mathcal{A}_{\text{score}}$) of the ImageNet ViT.

	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	0.418	0.373	0.496	0.358	0.579	0.539	0.499
Female	-	0.169	0.343	0.399	0.368	0.504	0.258
Male	-	-	0.183	0.381	0.213	0.187	0.225
No Hats	-	-	-	0.448	0.577	0.422	0.513
No Glasses	-	-	-	-	0.529	0.385	0.440
No Beards	-	-	-	-	-	0.509	0.440
No Beard No Hats	-	-	-	-	-	-	0.368

Table 27. The full results for the alignment scores ($\mathcal{A}_{\text{score}}$) of the CLIP ViT.

	Female	Male	No Hats	No Glasses	No Beards	No Beard No Hats	No Glasses No Smiles No Ties
Full CelebA	0.513	0.385	0.439	0.294	0.469	0.417	0.255
Female	-	0.203	0.456	0.313	0.312	0.521	0.342
Male	-	-	0.371	0.286	0.238	0.201	0.306
No Hats	-	-	-	0.322	0.411	0.378	0.302
No Glasses	-	-	-	-	0.459	0.427	0.286
No Beards	-	-	-	-	-	0.301	0.324
No Beard No Hats	-	-	-	-	-	-	0.236

Table 28. The full results for the alignment scores ($\mathcal{A}_{\text{score}}$) of the MAE ViT.