# Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints (Supplementary Material)

Ning Yu[1,2] Larry Davis[1] Mario Fritz[3]
[1]University of Maryland, College Park
[2]Max Planck Institute for Informatics
Saarland Informatics Campus, Germany
[3]CISPA Helmholtz Center for Information Security
Saarland Informatics Campus, Germany
ningyu@mpi-inf.mpg.de lsd@cs.umd.edu fritz@cispa.saarland

## 1. Fréchet Distance ratio

As described in Section 4.1 in the main paper, we use the ratio of inter-class and intra-class Fréchet Distance [3], denoted as FD ratio, to evaluate the distinguishability of a feature representation across classes. For inter-class FD calculation, we first measure the FD between two feature distributions from a pair of different classes, and then average over each possible pair. For intra-class FD calculation, we first measure the FD between two feature distributions from two disjoint sets of images in the same class, where we split the class equally, and then average over each class.

Mathematically,

$$\text{FD ratio} = \frac{\text{inter-class FD}}{\text{intra-class FD}} \tag{1}$$

$$\text{inter-class FD} = \frac{1}{||\{(y,\tilde{y})|y \neq \tilde{y}\}||} \sum_{y \neq \tilde{y}} \text{FD}\Big(\{f(I_i)|y_i = y\}, \{f(I_j)|y_j = \tilde{y}\}\Big) \tag{2}$$

$$\text{intra-class FD} = \frac{1}{||\mathbb{Y}||} \sum_{y \in \mathbb{Y}, \{i\} \cap \{j\} = \emptyset} \text{FD}\Big(\{f(I_i)|y_i = y\}, \{f(I_j)|y_j = y\}\Big) \tag{3}$$

where $\mathbb{Y}$ is the class set for image sources and $f(\cdot)$ is a feature representation mapping from image domain to a feature domain.

Then in all the tables in the main paper, we compare FD ratio between the inception feature [8] as a baseline and our learned features. The larger the ratio, the more distinguishable the feature representation across sources. We also show in Figure 1 in the main paper the t-sne visualization [6] of the two features.

## 2. Face samples

We show more face samples corresponding to the experiments in the main paper. See Figure 1 to 14.

## 3. Bedroom samples

We show bedroom samples corresponding to the experiments in the main paper. See Figure 15 to 29. In general, LSUN bedroom dataset is more challenging to a GAN model because of lack of image alignment. However, ProGAN [4] still performs equally well on this dataset and does not affect our conclusions in the main paper.

## References

[1] Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017. 5, 19

[2] Mikoaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. 6, 20

[3] DC Dowson and BV Landau. The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982. 1

[4] 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, 17

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 2

[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 1

Figure 1. Face samples from CelebA real dataset [5]


[7] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 4, 18

[8] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2226–2234, 2016. 1

[9] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 16

[10] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 8, 22

Figure 2. Face samples from ProGAN [4]

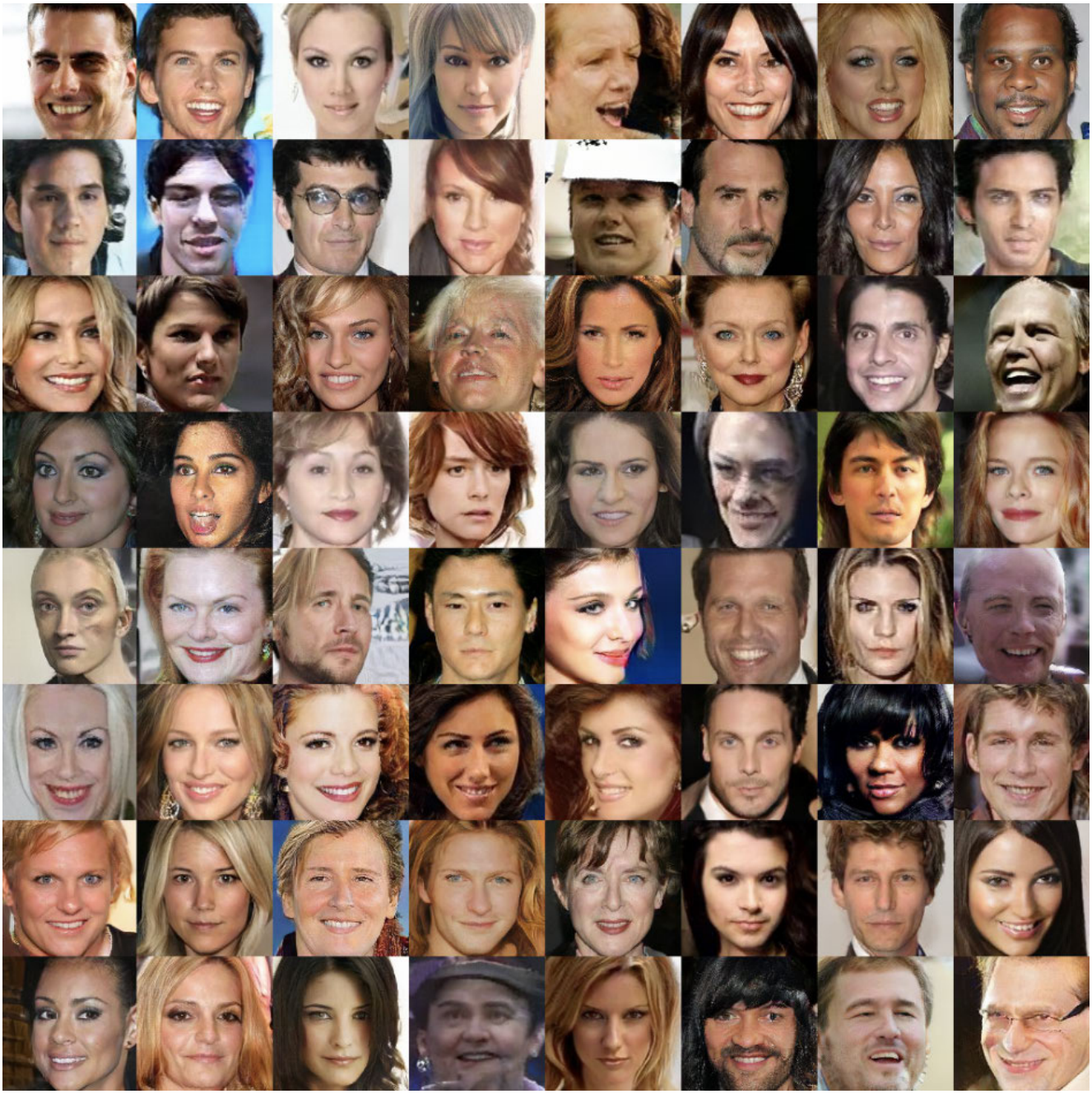

Figure 3. Face samples from SNGAN [7]

Figure 4. Face samples from CramerGAN [1]

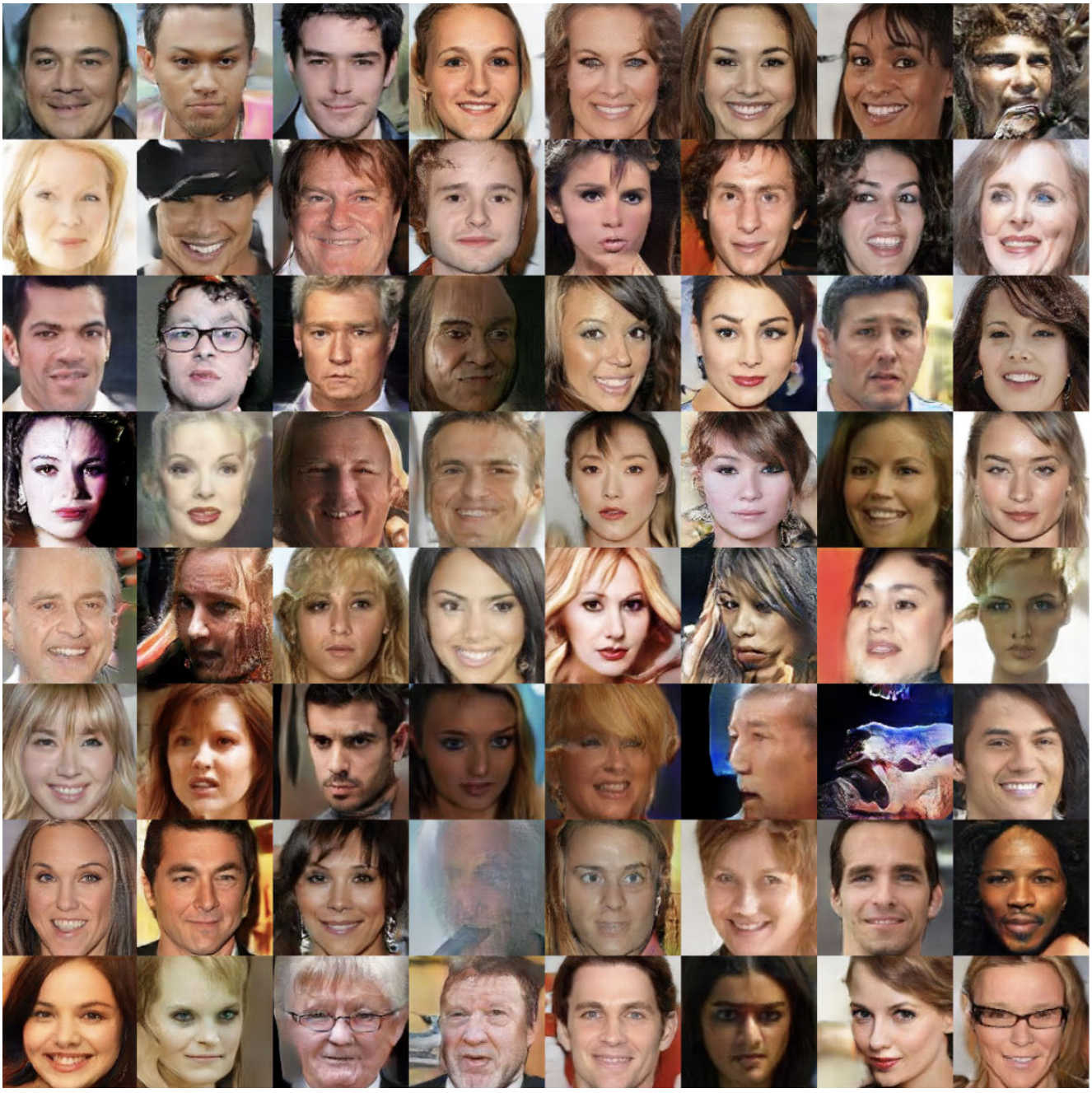

Figure 5. Face samples from MMDGAN [2]

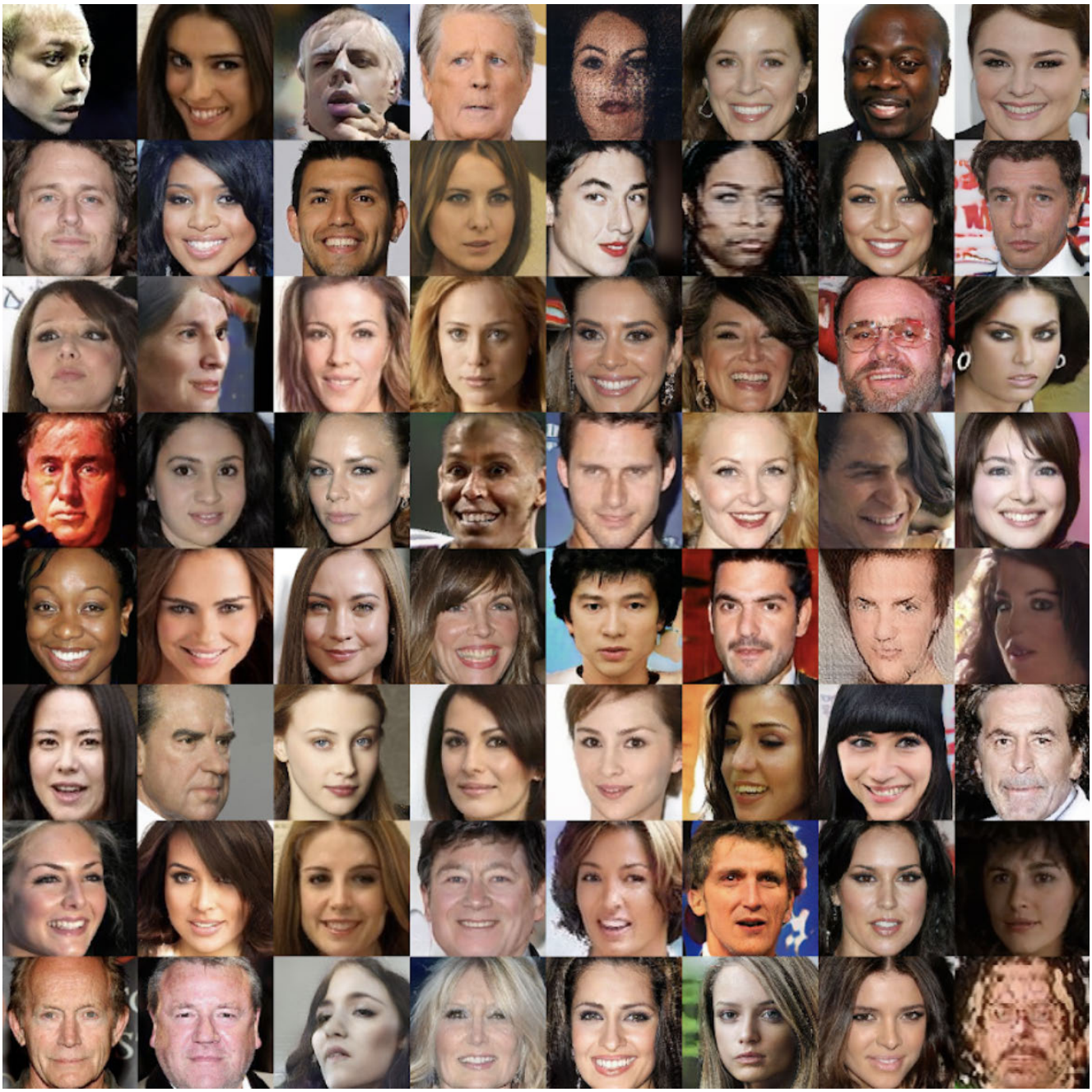

Figure 6. Arbitrary face samples from the setup of {*real*, *ProGAN_seed_v#i*} where $i \in \{1, ..., 10\}$.

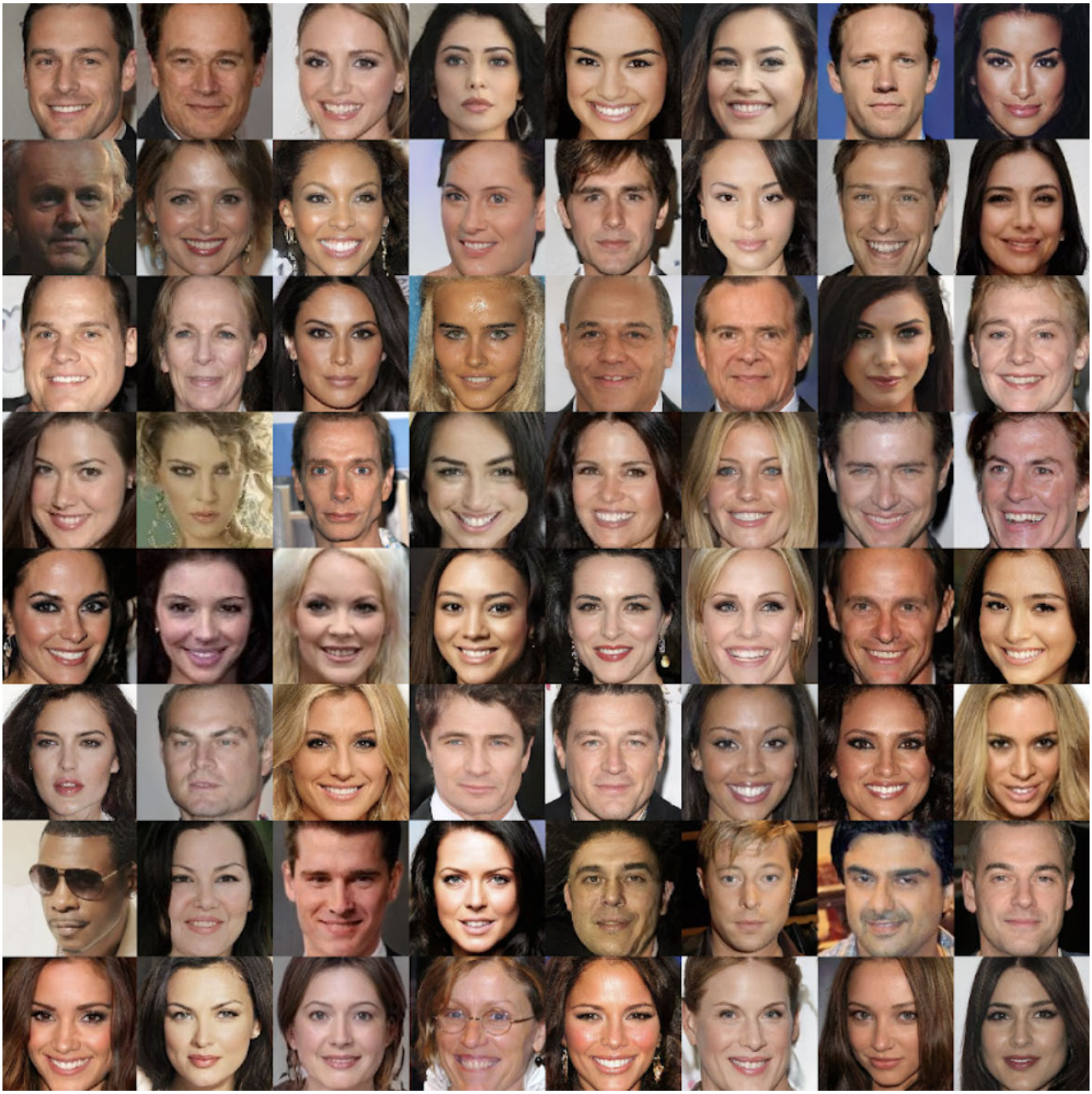

Figure 7. Filtered face samples from the setup of {*real*, *ProGAN_seed_v#i*} with the top 10% largest Perceptual Similarity [10] to real dataset distribution.

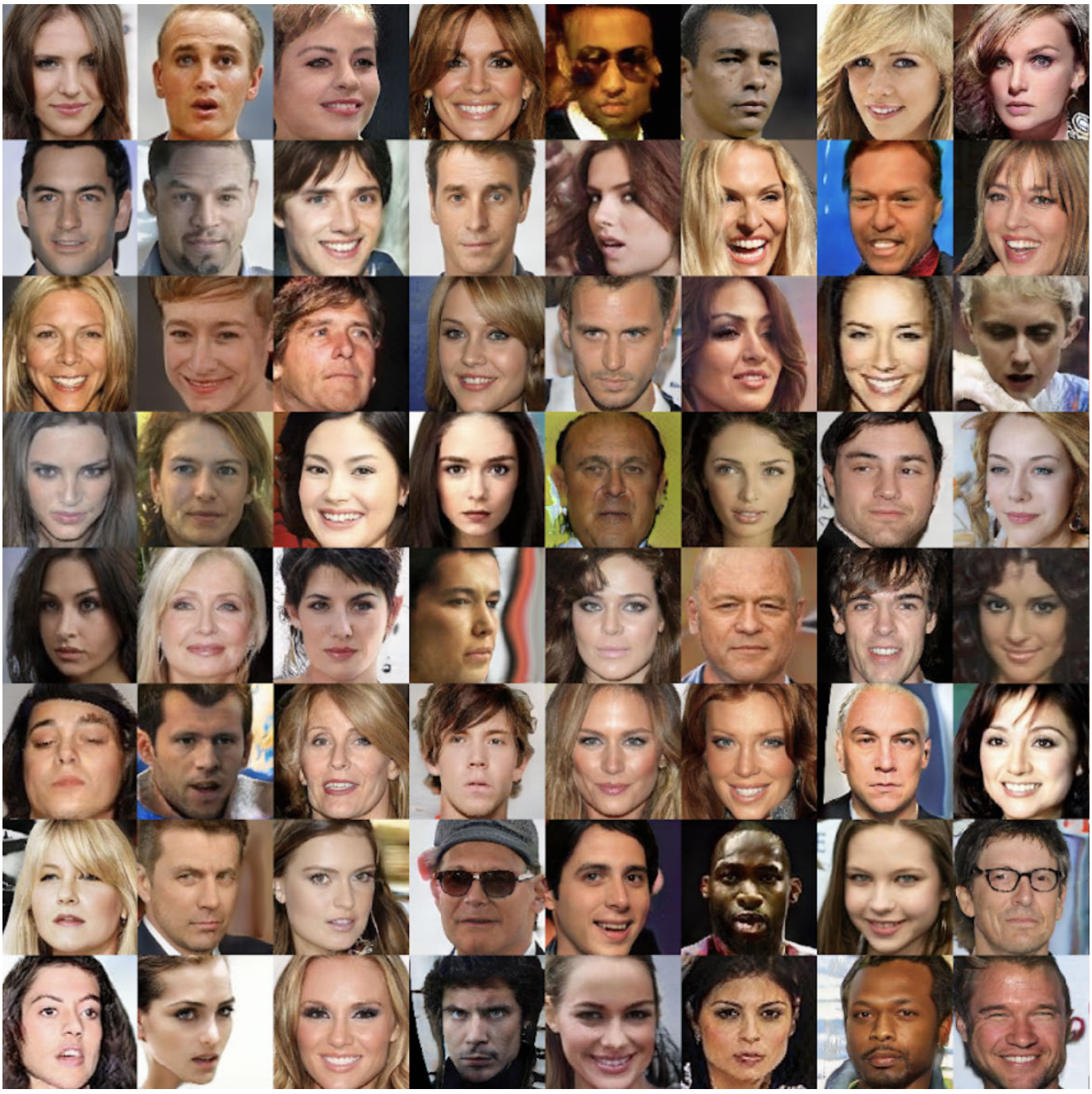

Figure 8. Arbitrary face samples without attack from the setup of {*real*, *ProGAN_seed_v#i*}.

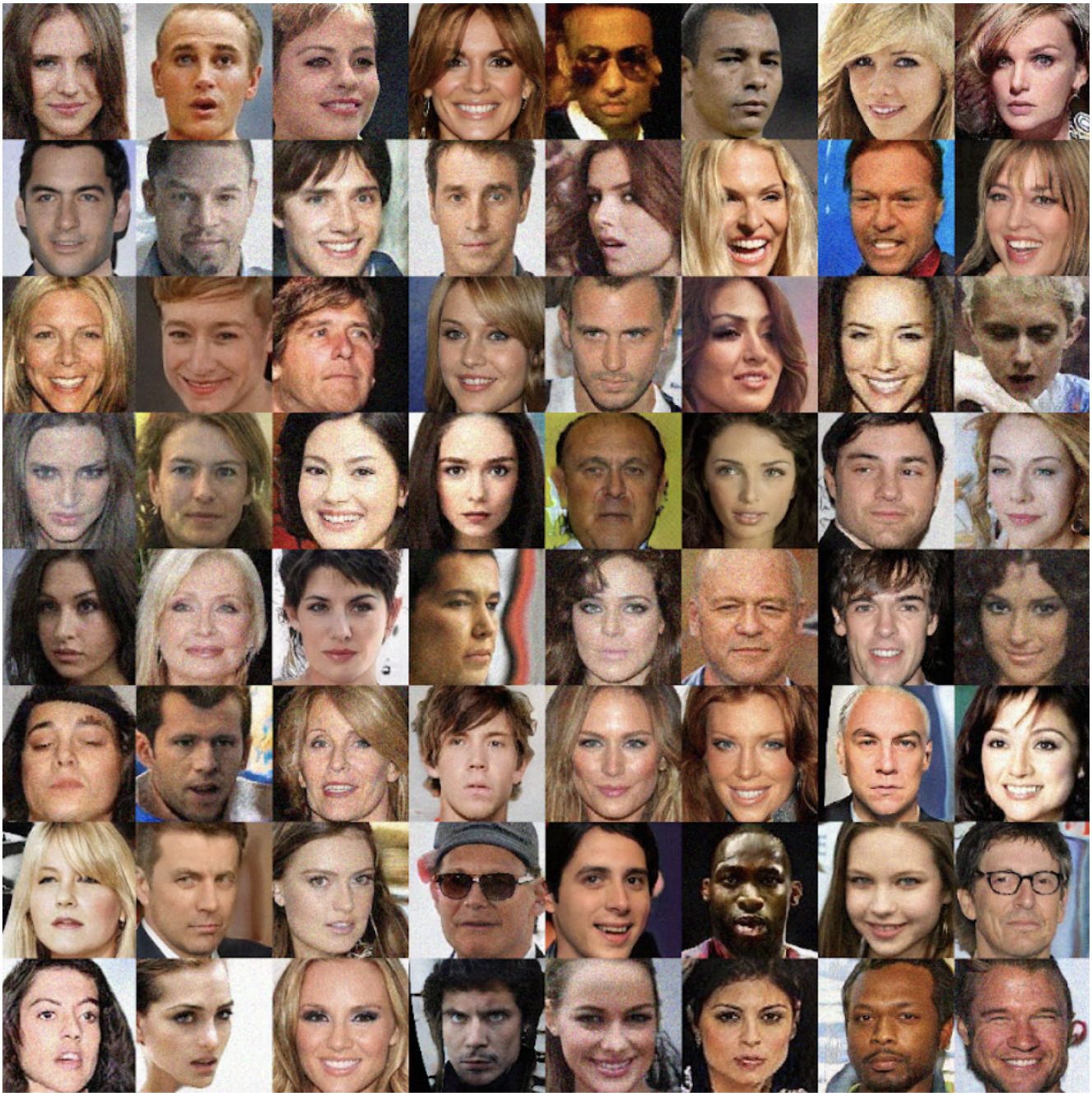

Figure 9. Arbitrary face samples with *noise* attack from the setup of {*real*, *ProGAN_seed_v#i*}.

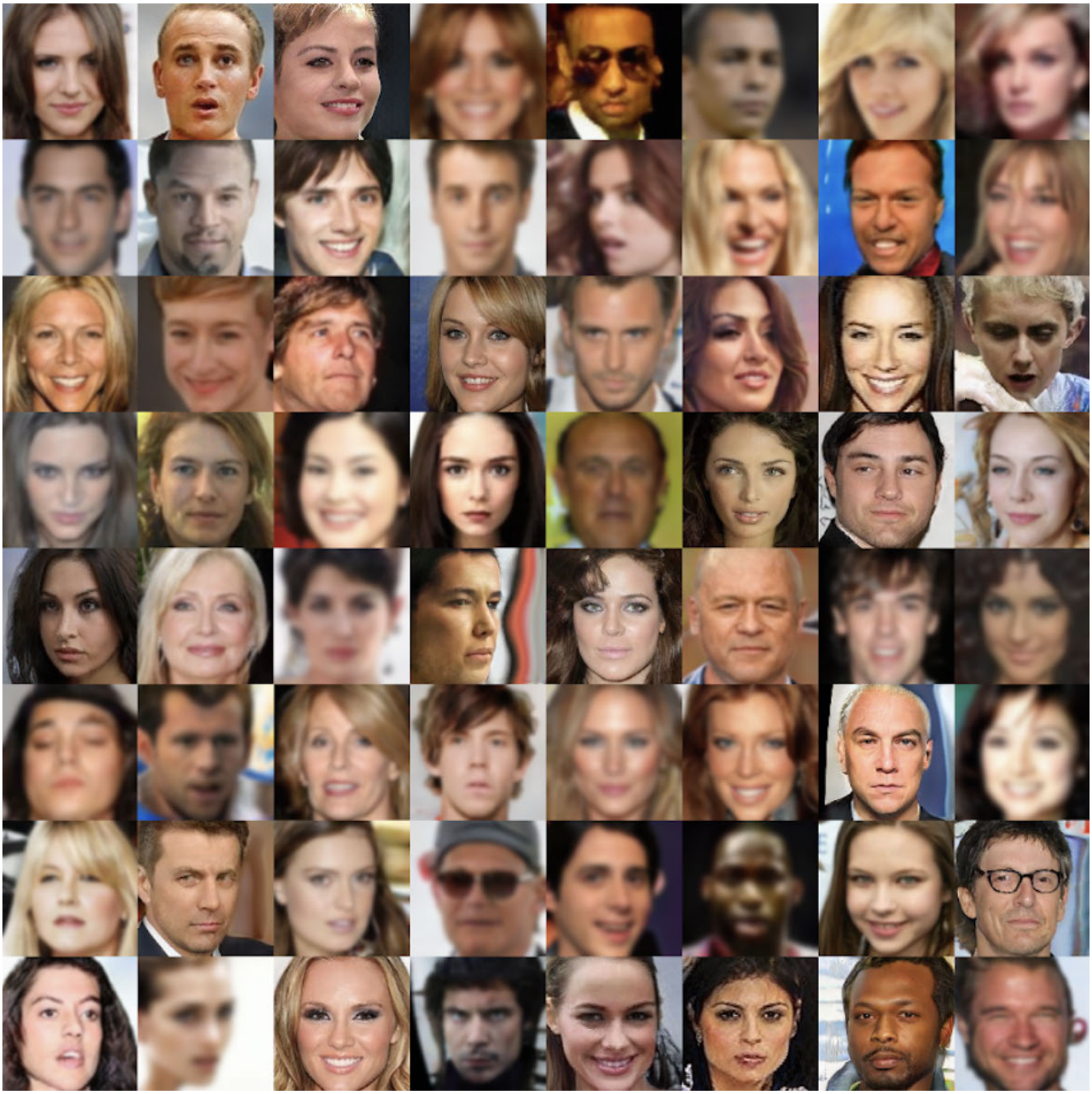

Figure 10. Arbitrary face samples with *blur* attack from the setup of {*real*, *ProGAN_seed_v#i*}.

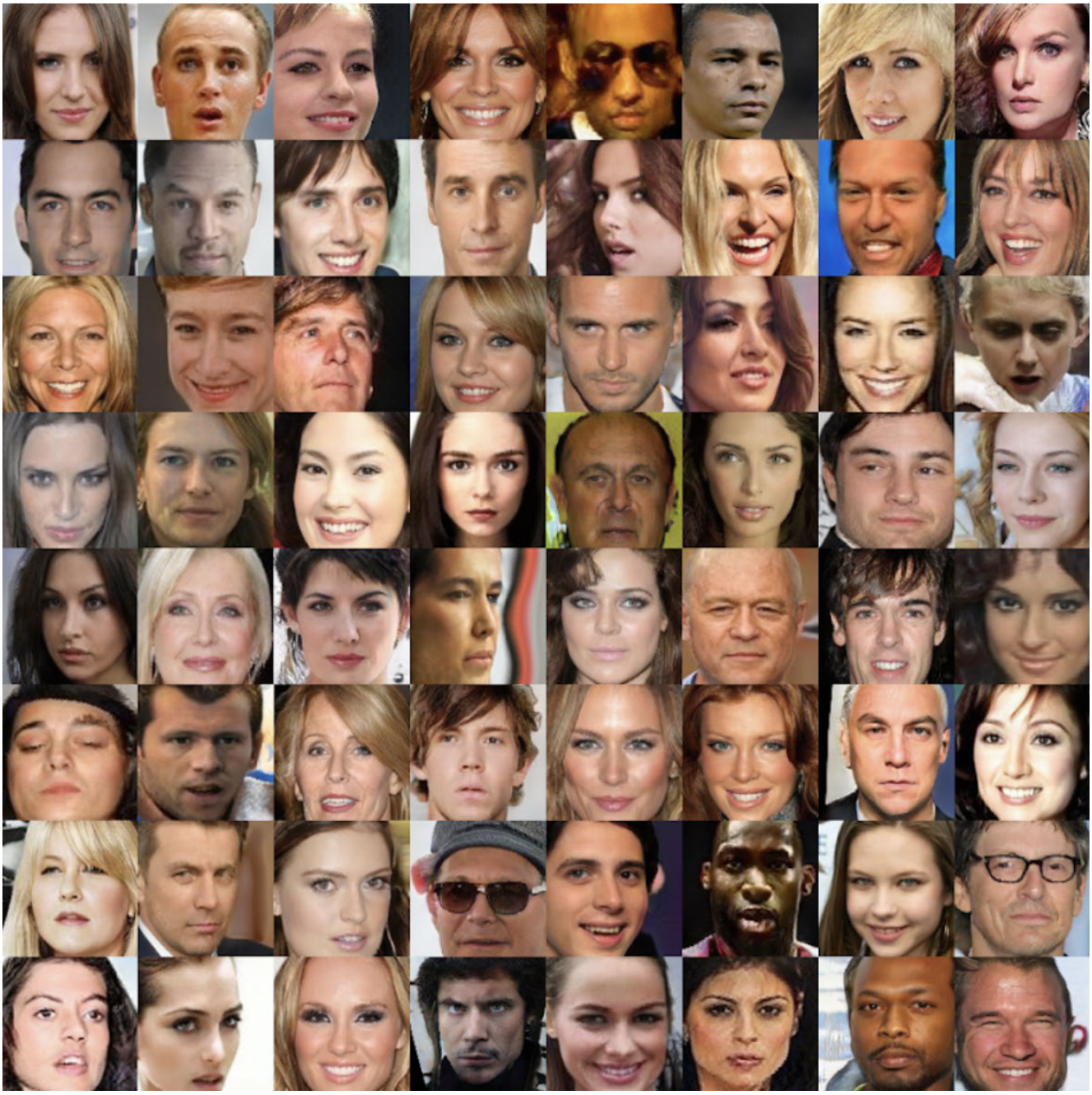

Figure 11. Arbitrary face samples with *cropping* attack from the setup of {*real*, *ProGAN_seed_v#i*}.

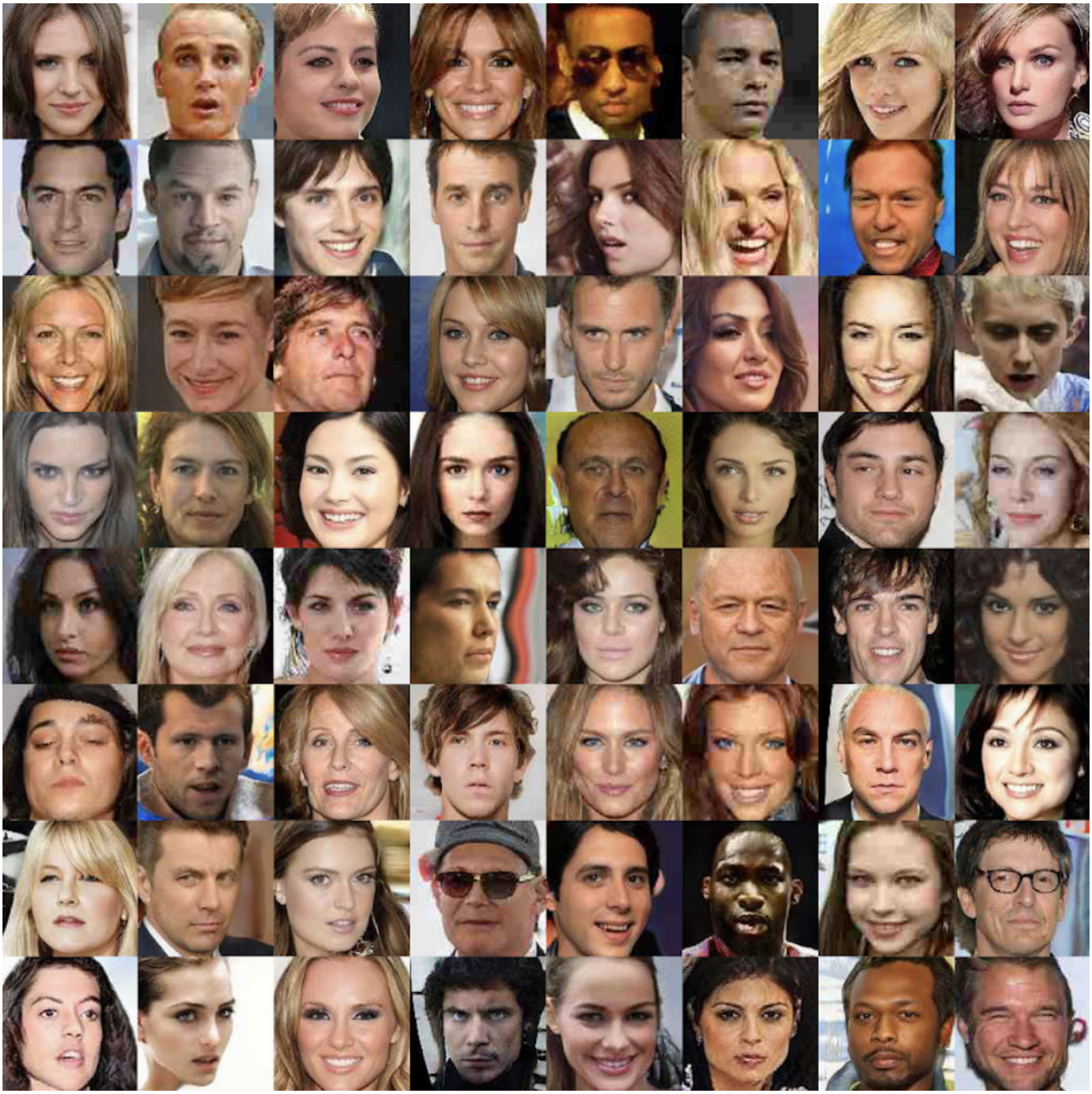

Figure 12. Arbitrary face samples with *JPEG compression* attack from the setup of {*real*, *ProGAN_seed_v#i*}.

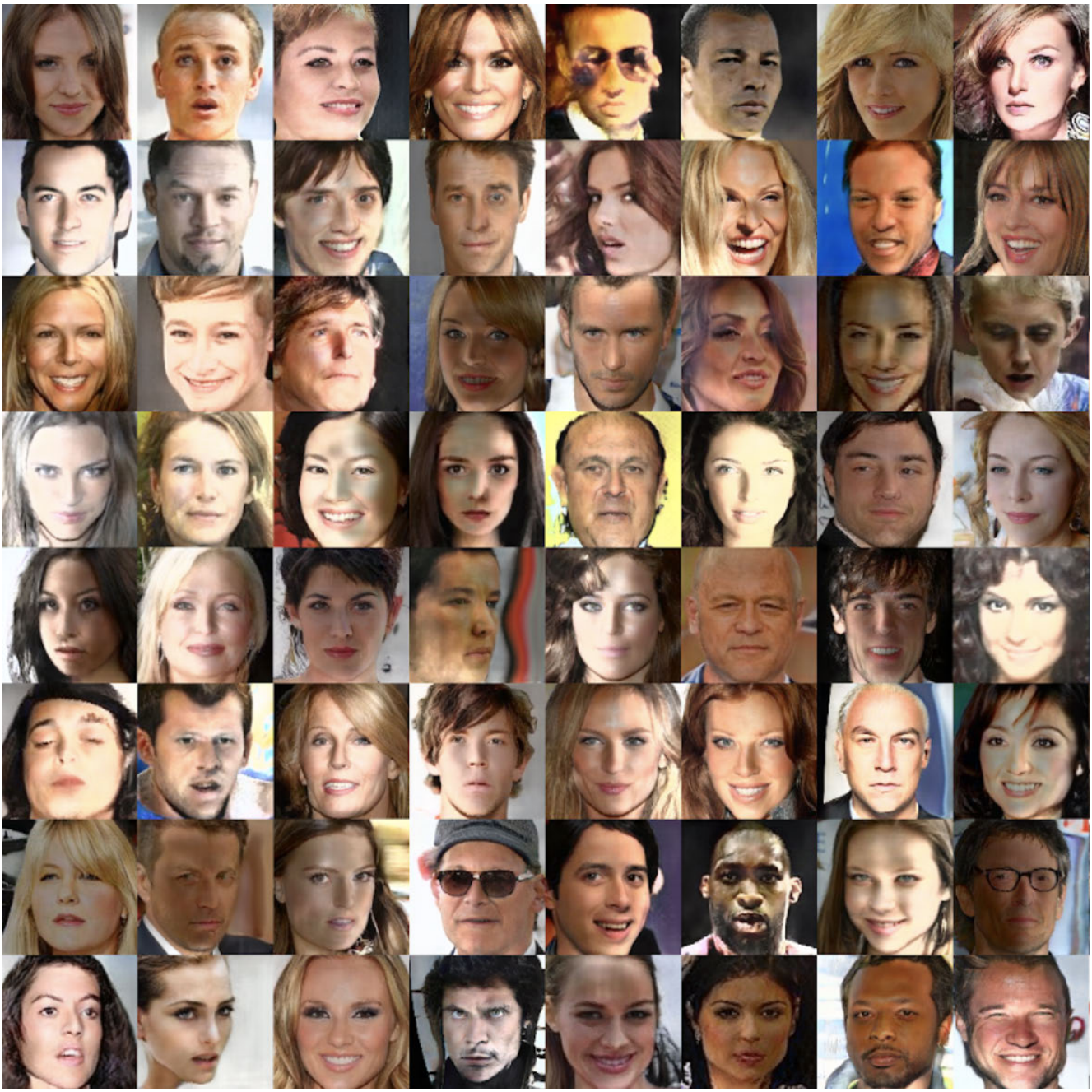

Figure 13. Arbitrary face samples with *relighting* attack from the setup of {*real*, *ProGAN_seed_v#i*}.

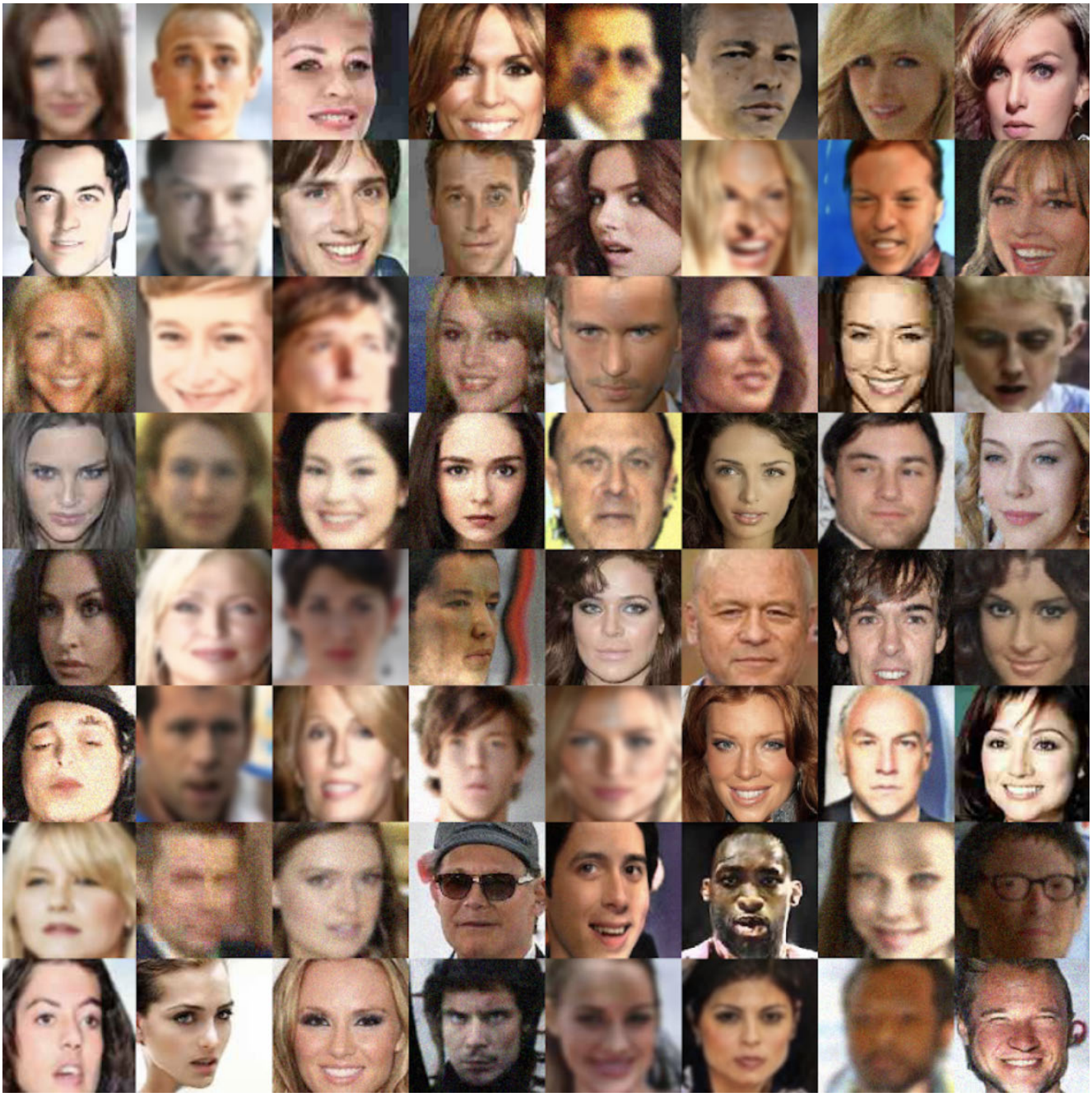

Figure 14. Arbitrary face samples with the combination attack from the setup of {*real*, *ProGAN_seed_v#i*}.

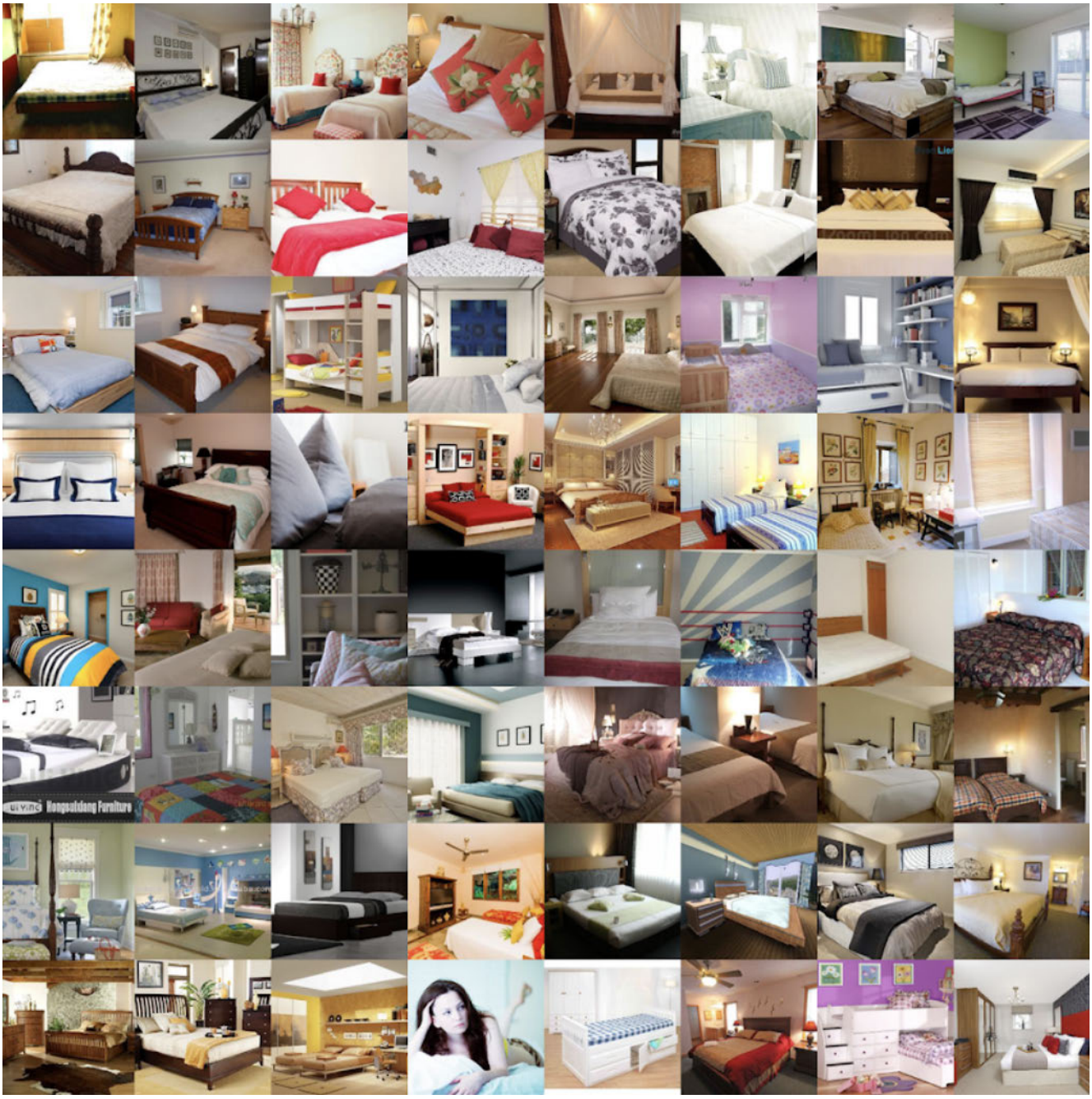

Figure 15. Bedroom samples from LSUN real dataset [9]

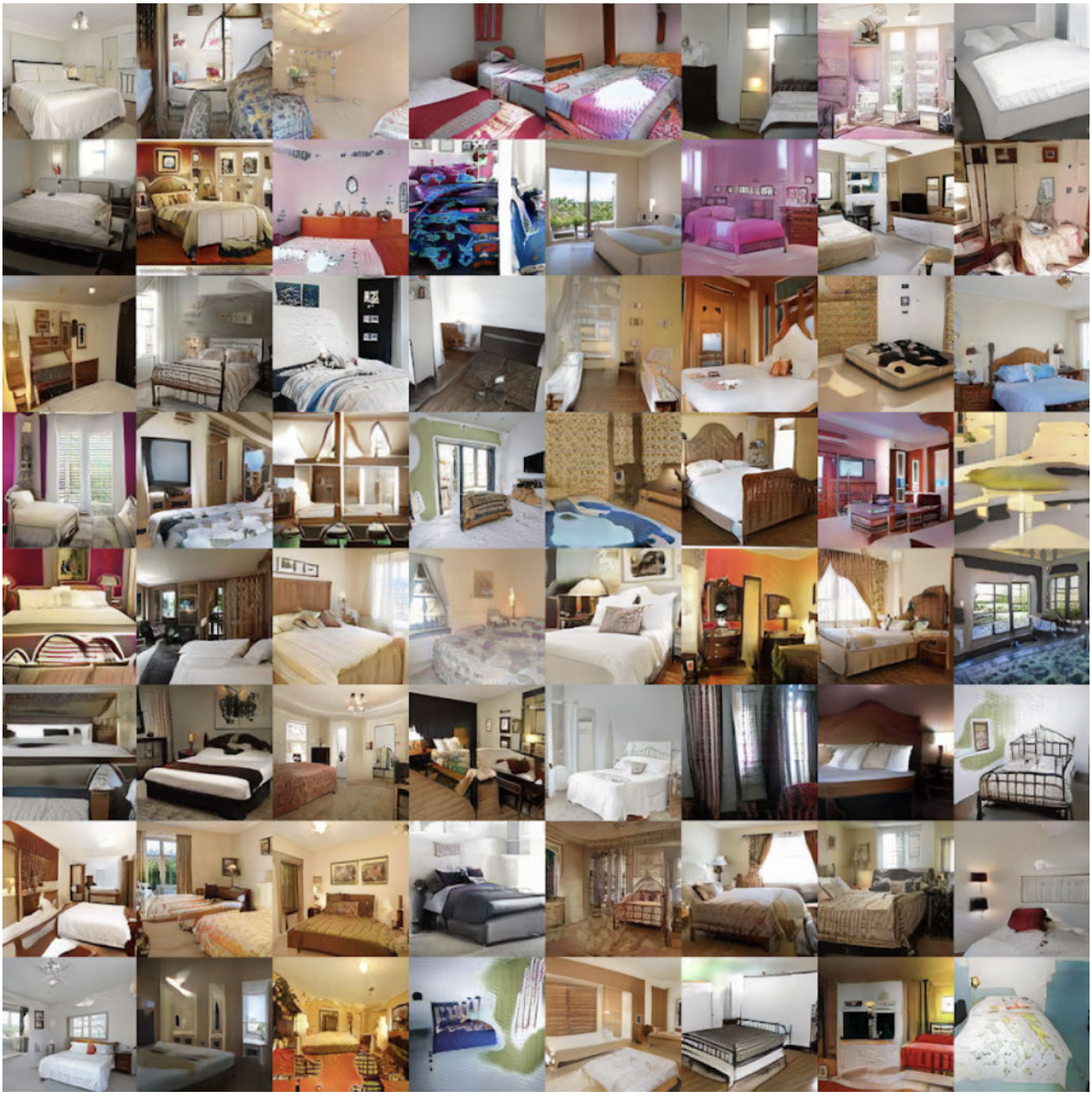

Figure 16. Bedroom samples from ProGAN [4]

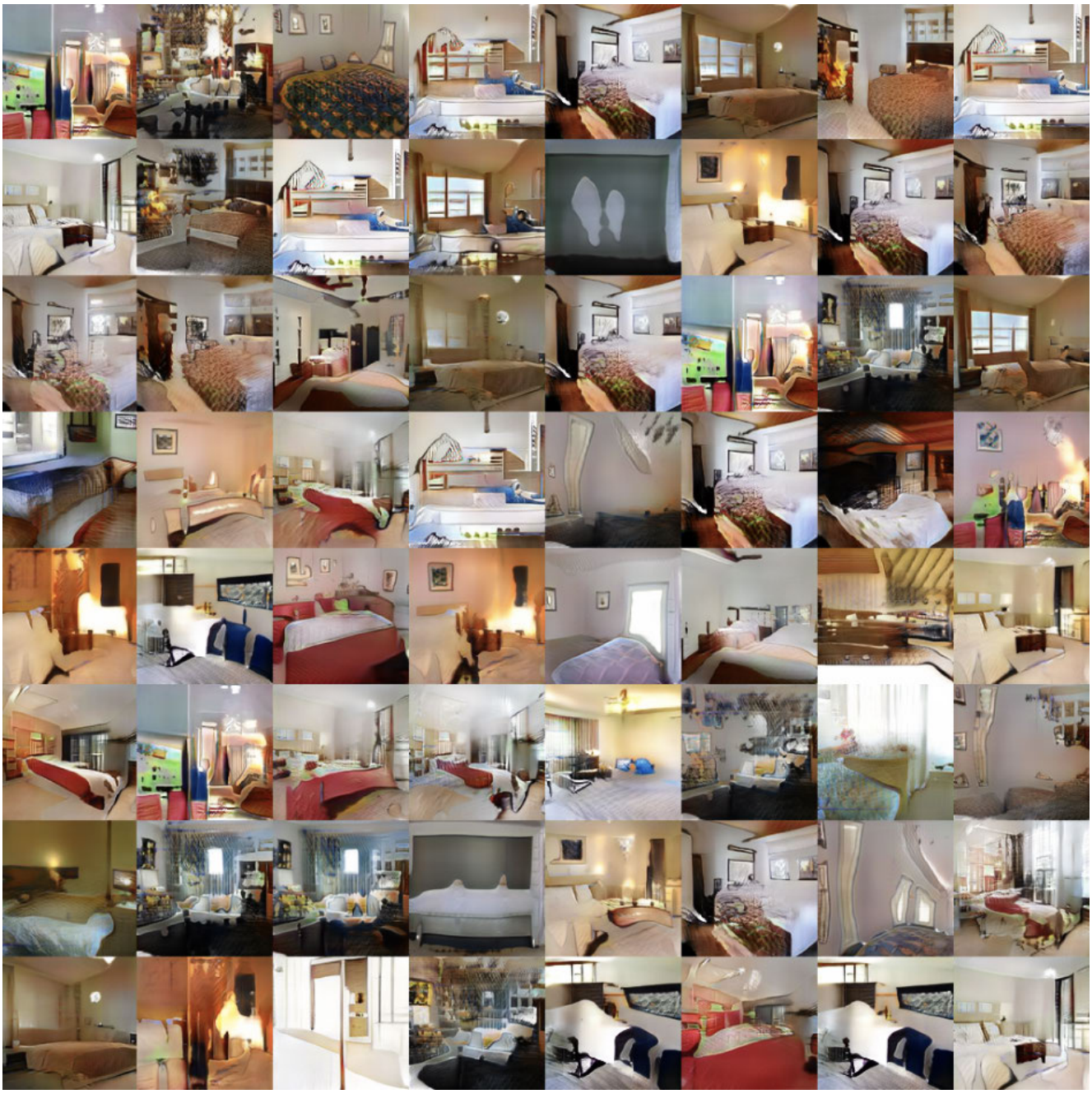

Figure 17. Bedroom samples from SNGAN [7]

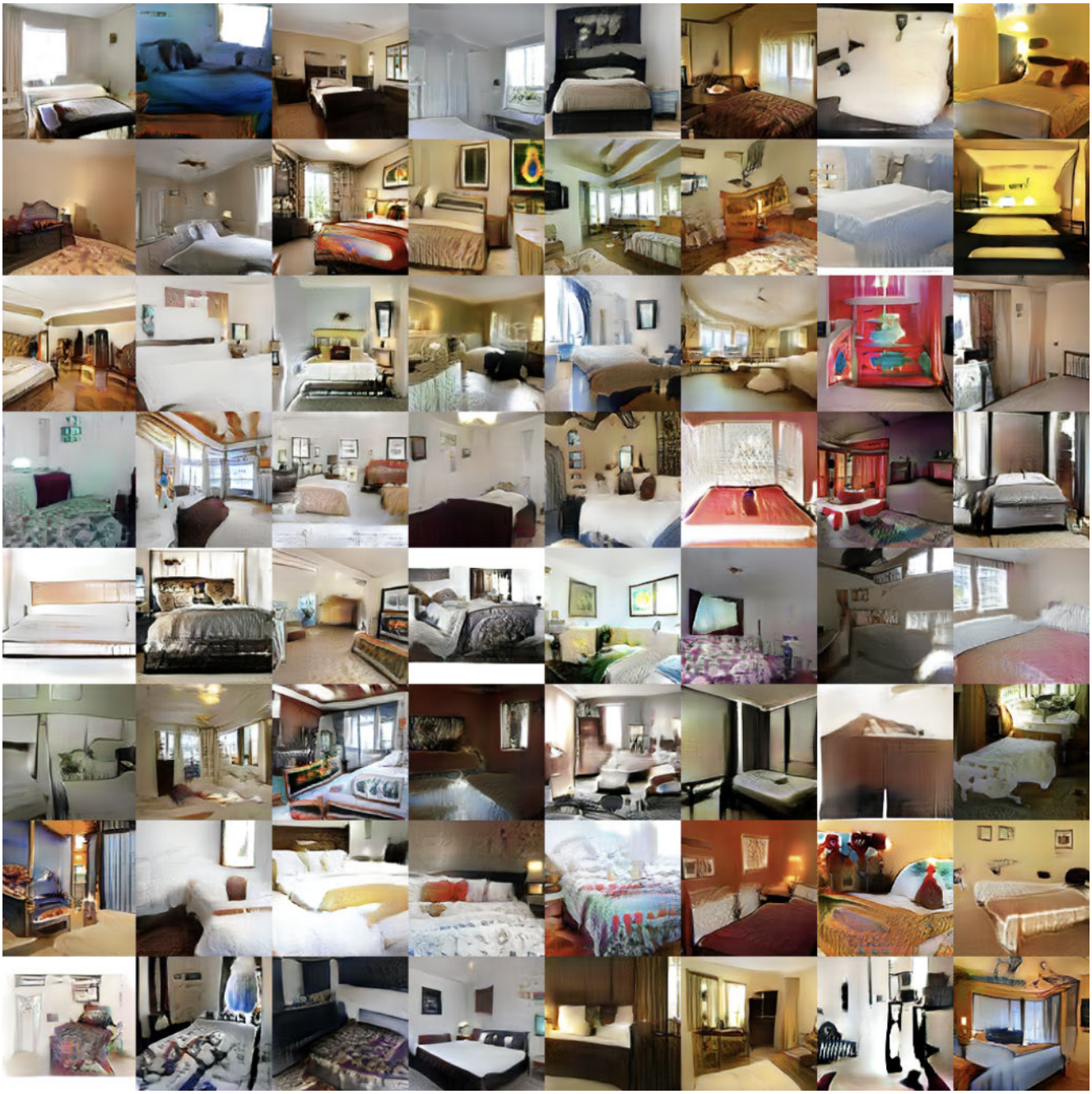

Figure 18. Bedroom samples from CramerGAN [1]

Figure 19. Bedroom samples from MMDGAN [2]

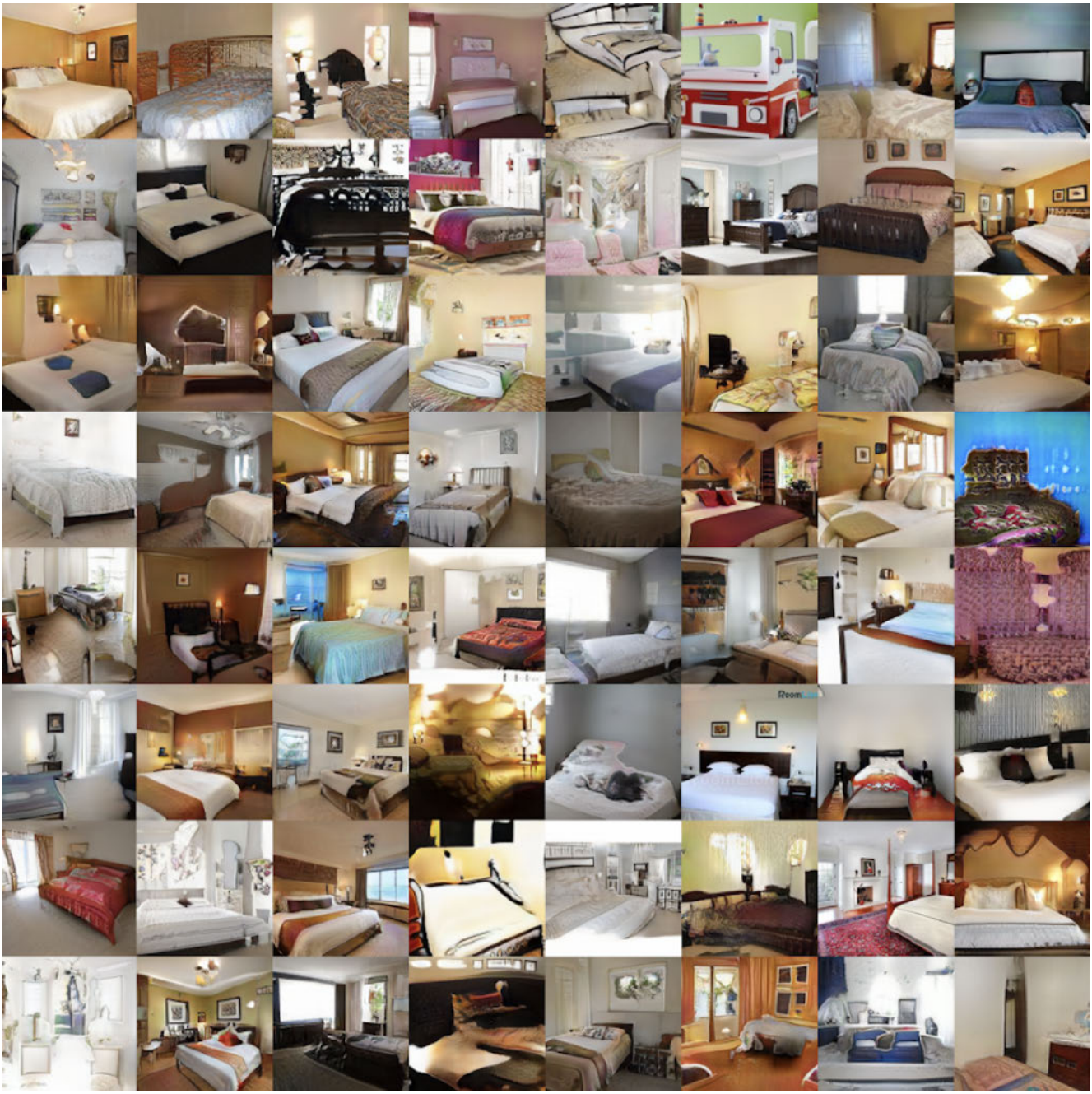

Figure 20. Arbitrary bedroom samples from the setup of {*real*, *ProGAN_seed_v#i*} where $i \in \{1, ..., 10\}$.

Figure 21. Filtered bedroom samples from the setup of {*real*, *ProGAN_seed_v#i*} with the top 10% largest Perceptual Similarity [10] to real dataset distribution.

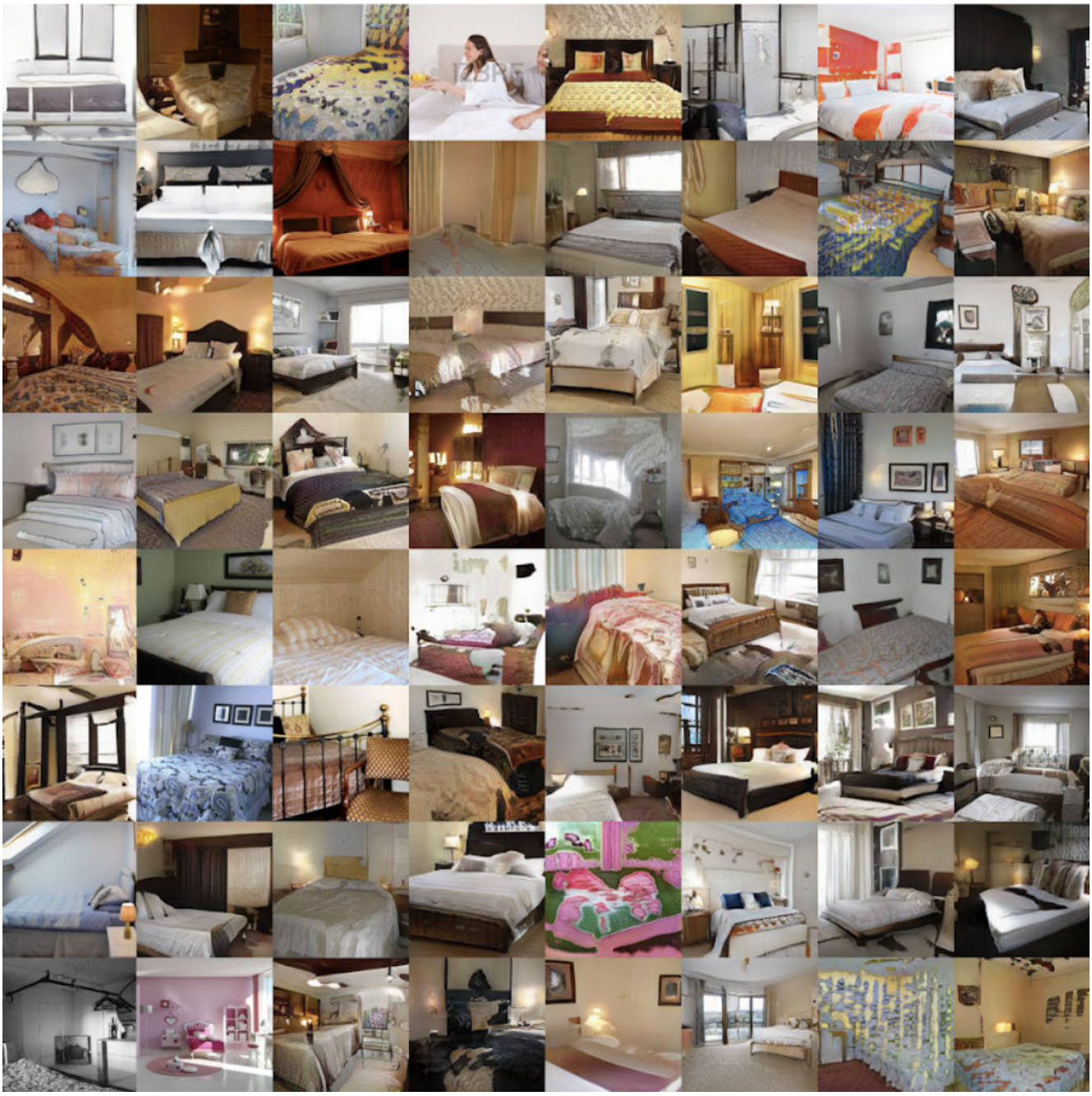

Figure 22. Arbitrary bedroom samples without attack from the setup of {*real*, *ProGAN_seed_v#i*}.

Figure 23. Arbitrary bedroom samples with *noise* attack from the setup of {*real*, *ProGAN_seed_v#i*}.

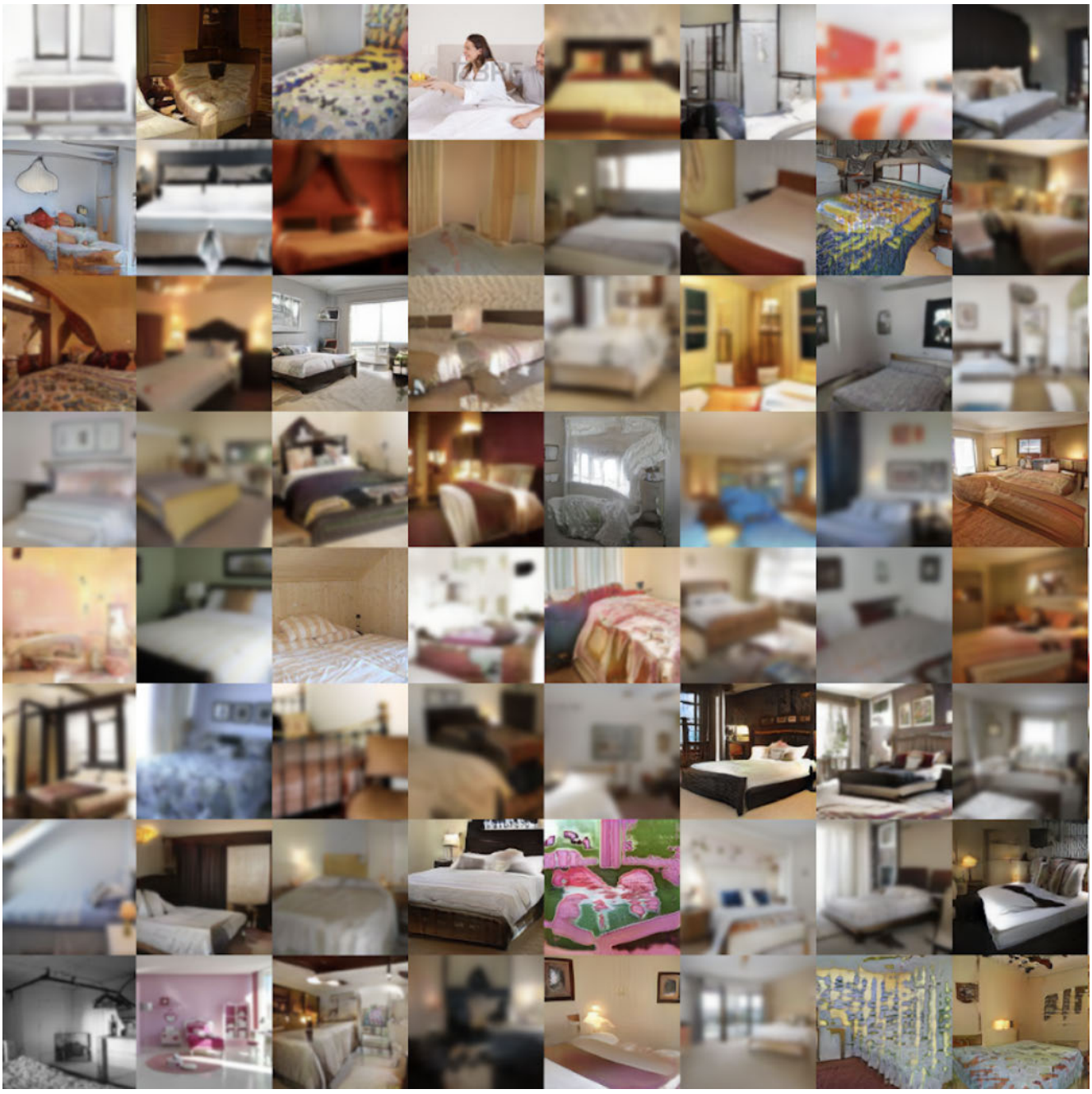

Figure 24. Arbitrary bedroom samples with *blur* attack from the setup of {*real*, *ProGAN_seed_v#i*}.

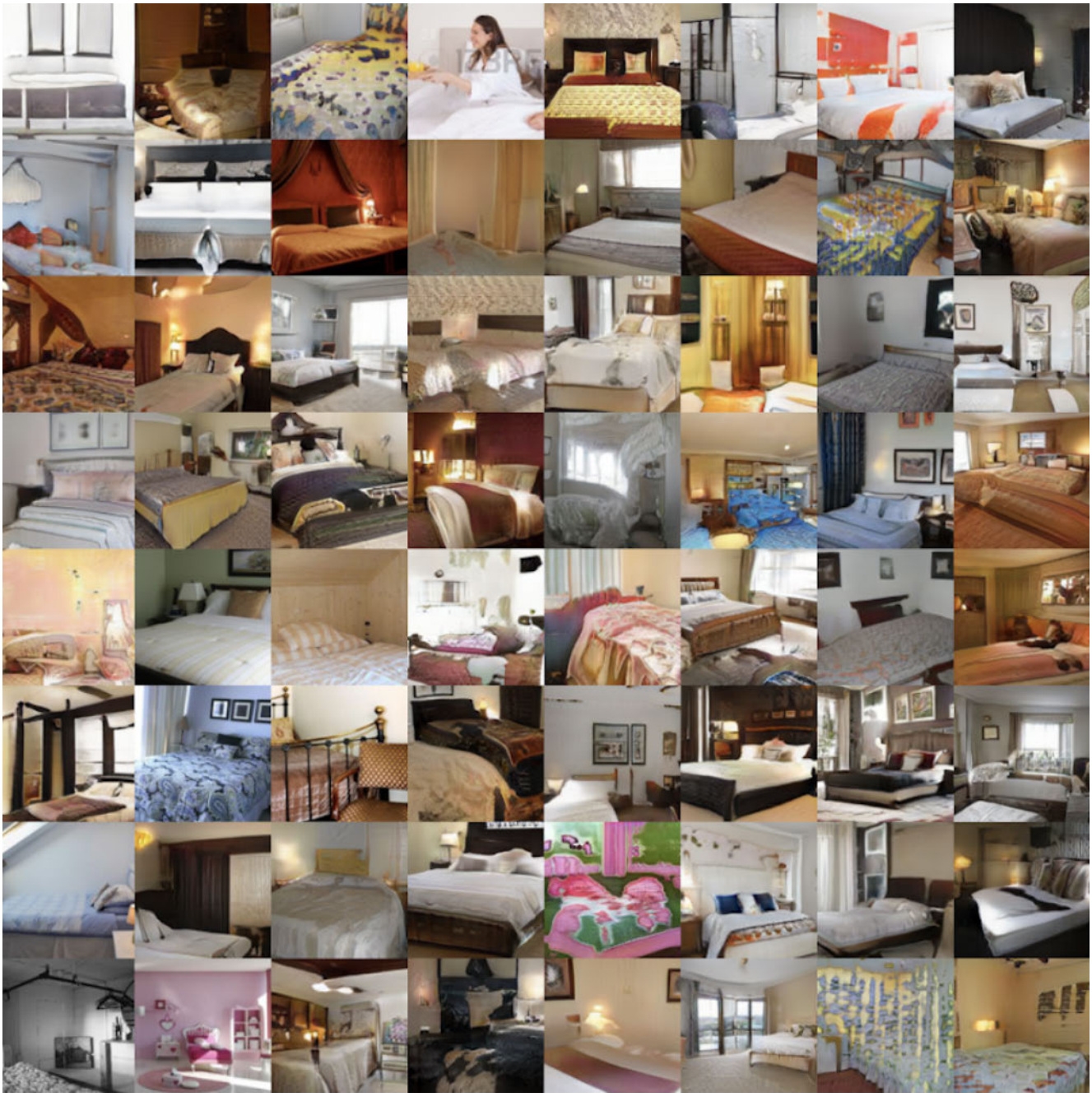

Figure 25. Arbitrary bedroom samples with *cropping* attack from the setup of {*real*, *ProGAN_seed_v#i*}.

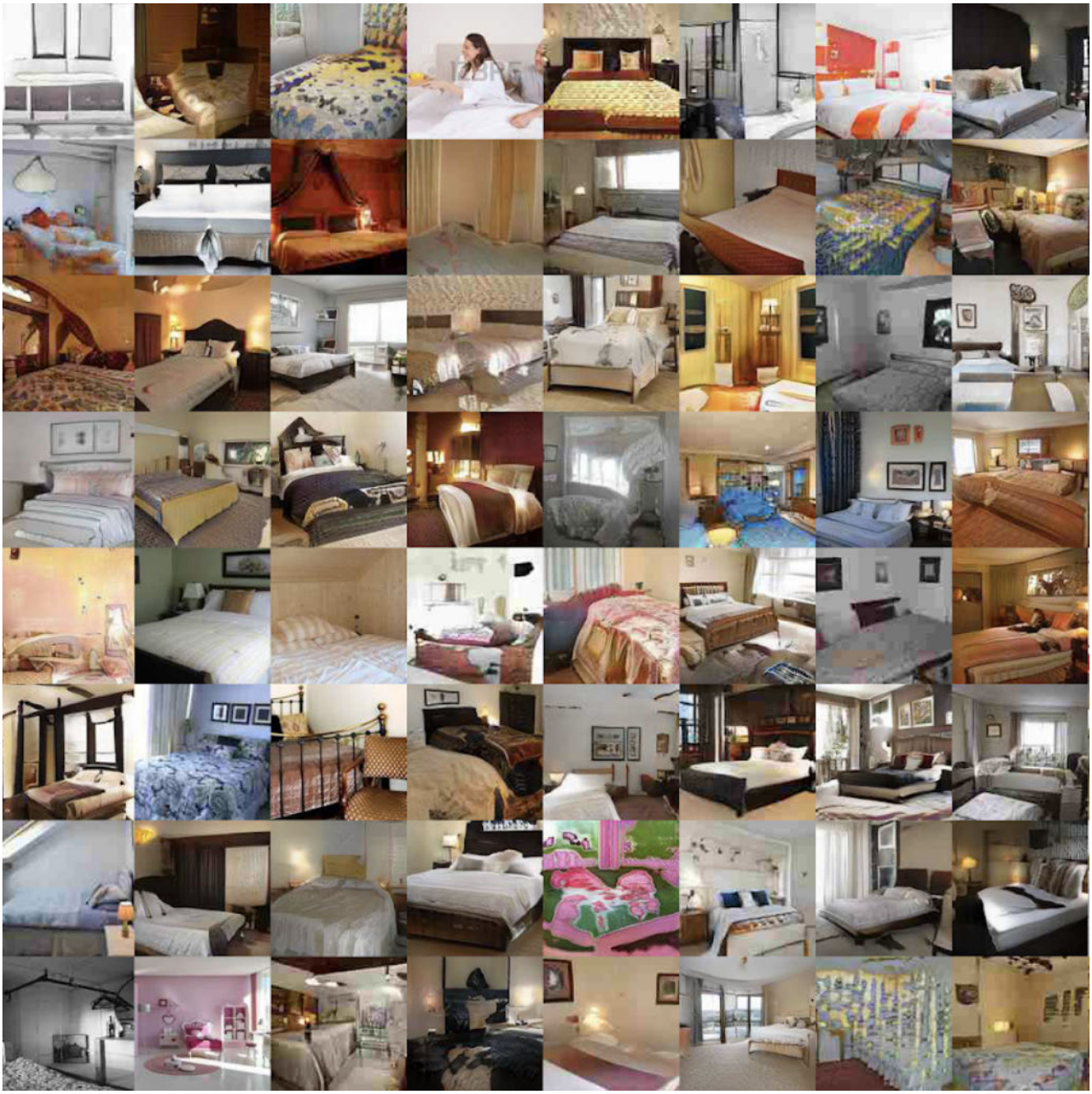

Figure 26. Arbitrary bedroom samples with *JPEG compression* attack from the setup of {*real*, *ProGAN_seed_v#i*}.

Figure 27. Arbitrary bedroom samples with *relighting* attack from the setup of {*real*, *ProGAN_seed_v#i*}.

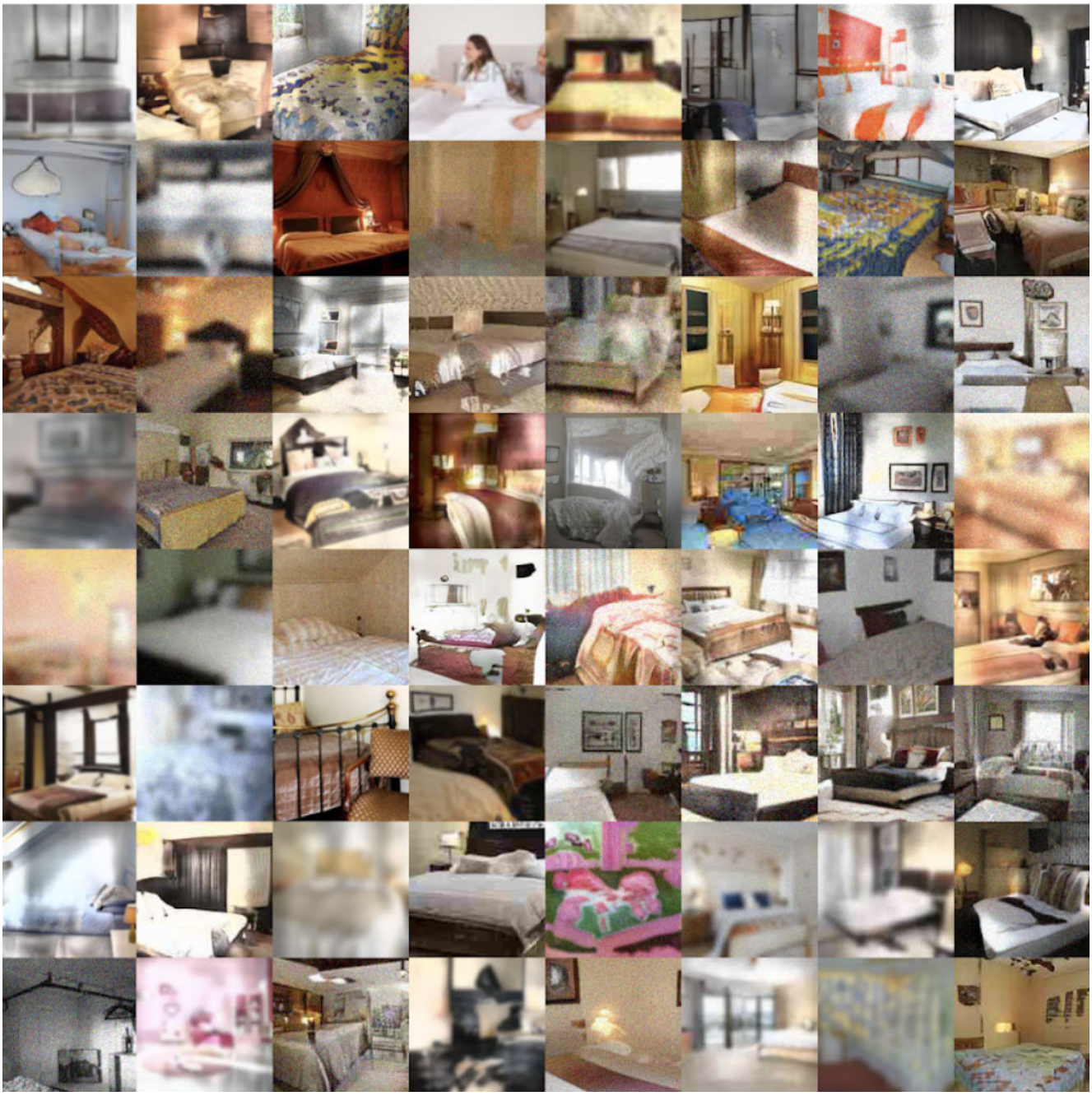

Figure 28. Arbitrary bedroom samples with the combination attack from the setup of {*real*, *ProGAN_seed_v#i*}.

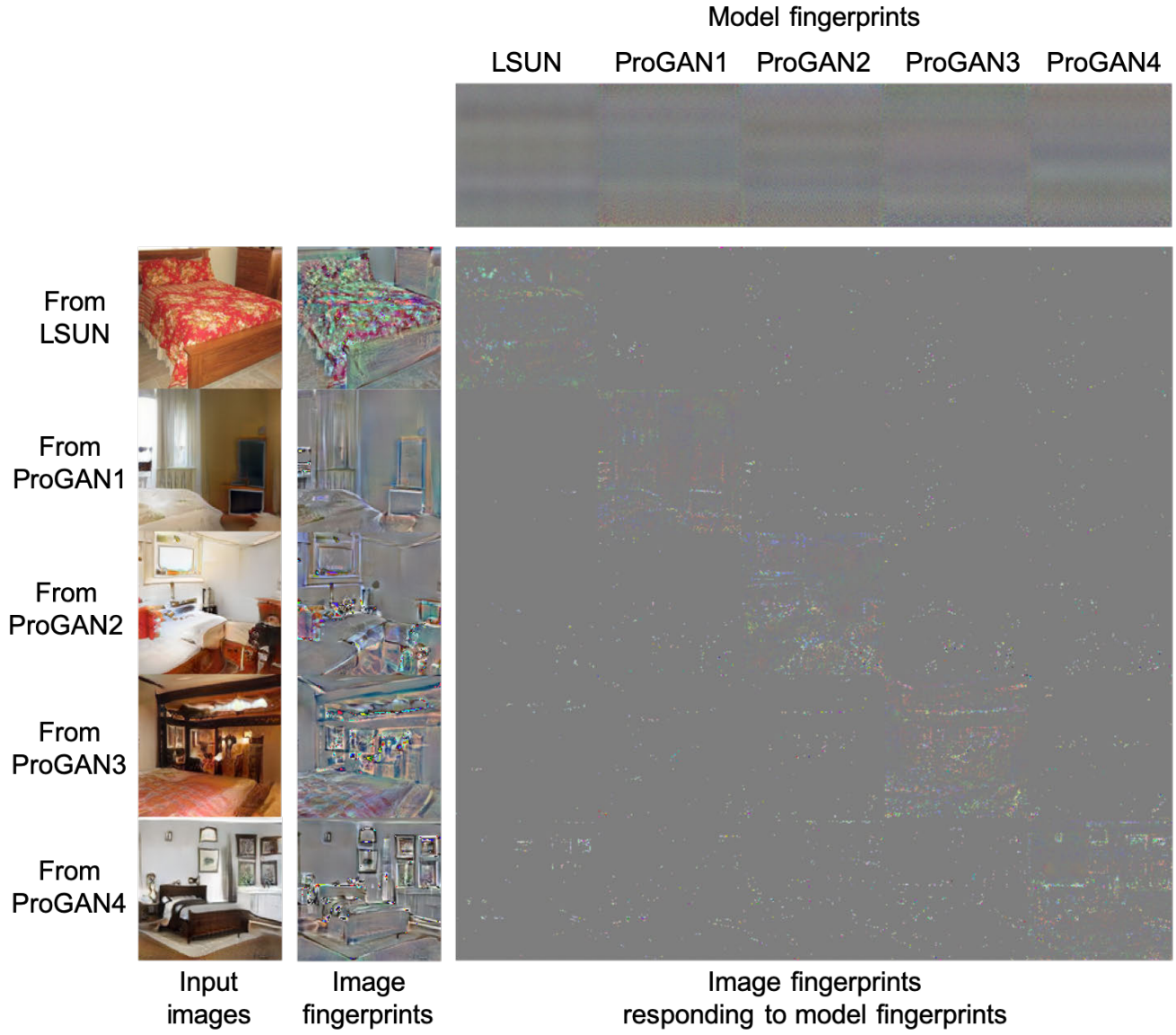


Figure 29. Visualization of bedroom model and image fingerprint samples. Their pairwise interactions are shown as the confusion matrix. It turns out that image fingerprints maximize responses only to their own model fingerprints, which supports effective attribution.