

Supplemental material:

GANmut: Learning Interpretable Conditional Space for Gamut of Emotions

Stefano d’Apolito¹, Danda Pani Paudel¹, Zhiwu Huang¹, Andrés Romero¹, Luc Van Gool^{1,2}

¹Computer Vision Lab, ETH Zürich, Switzerland ²PSI, KU Leuven, Belgium

dstefano@alumni.ethz.ch {paudel, zhiwu.huang, roandres, vangool}@vision.ee.ethz.ch

1. Structure of this document

In the document we first provide the details of our method, which were missing due to the lack of space in the main paper. Later we will provide more qualitative and some quantitative results for further analysis. A section that continues the discussion of the paper is also introduced.

2. Gaussian model loss

Besides losses \mathcal{L}_{cls}^f (eq. 10) and \mathcal{L}_{div} (eq. 11), to complete the loss terms of the Gaussian model we have:

$$\mathcal{L}_{coord} = \mathbb{E}_{x,z} \left[\|D_{coord}(G(x,z)) - z\|_2^2 \right] \quad (12)$$

$$\mathcal{L}_{rec} = \mathbb{E}_{x,z} [\|x - G(G(x,z), D_{coord}(x))\|_1] \quad (13)$$

$$\mathcal{L}_{cls}^r = \mathbb{E}_{x,c'} [-\log D_{cls}(c' | x)] \quad (14)$$

$$\begin{aligned} \mathcal{L}_{adv} = & \mathbb{E}_x [D_{src}(x)] - \mathbb{E}_{x,c} [D_{src}(G(x,c))] \\ & - \lambda_{gp} \mathbb{E}_{\hat{x}} \left[(\|\nabla_{\hat{x}} D_{src}(\hat{x})\|_2 - 1)^2 \right] \end{aligned} \quad (15)$$

Therefore, the final loss is:

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls-D} \mathcal{L}_{cls}^r + \lambda_{cor} * \mathcal{L}_{cor} \quad (16)$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls-G} \mathcal{L}_{cls}^f + \lambda_{rec} \mathcal{L}_{rec} + \lambda_{div} * \mathcal{L}_{div} + \lambda_{cls} * \mathcal{L}_{cls} \quad (17)$$

3. More implementation and training details

In this section we will first give more details about the implementation of the linear model and then we will pass to the experimental evaluations.

GANmut. The learnable vector θ_c was encoded through a unitary L2 vector $v_c = \frac{[v_{c1}, v_{c2}]^t}{\|[v_{c1}, v_{c2}]^t\|_2}$, $v_{c1}, v_{c2} \in \mathbb{R}$ randomly initialized from $\mathcal{U}(-10^{-4}\sqrt{1/6}, 10^{-4}\sqrt{1/6})$. In

practice, samples from S_c were conditioned with code $\hat{z} = (\theta_c, \rho) = \rho * v_c$. Additionally, to make more robust the learning process, Gaussian noise of standard deviation equal to 0.1 was added to \hat{z} , before to pass it to the generator. **GGANmut.** $\mu_c = \tanh(w_c)$ was randomly initialized with default Pytorch settings, (i.e., w_c was sampled from $\mathcal{U}(\sqrt{0.5}, \sqrt{0.5})$). θ_c was initialized with 0, and $\sigma_{c,c}^2$ with 1. **Training.** The same training strategy of StarGAN was adopted: Adam with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, five discriminator updates every one of the generator, a batch size of 16 samples, a linear weight decay to 0 in the last 100K iterations and image augmentation. The GPUs farm used included Titan X, Titan Xp, GTX 1080 Ti. The numbers of iterations performed in 1 day varied between 130K and 250K, depending on the GPU and on the method.

4. Classification Error

As the authors of StarGAN evaluate quantitatively their method based on the classification accuracy obtained by the generated images, we replicate here the test. We again used the VGG trained on AffectNet. To obtain the results, for each emotion c we conditioned all the images of the test set. For GANmut we have taken the conditional code $z = (\theta_c, 1)$, for GGANmut $z = \mu_c$, and for SMIT we conditioned on c , sampling new random noise each time. Results are shown in table 1. As it can be seen, real images present expressions much more ambiguous than generated ones.

5. Why FED score for GGANmut is so high?

The left plot of Figure 1 may suggest why GGANmut, conditioned on emotion means μ_c (plus some small random noise, as in the evaluation settings) have a FED score much higher than the other methods. In particular, as one can notice, its VGG-features distribution clearly clusters in 7 groups. These clusters are likely to represent the 7 categorical emotions. Hence GGANmut manages, if conditioned on means, to produce intense expressions that are more easily classifiable than competitors. This is confirmed also by results in table 1. However, spontaneous expressions are of-

	Neutral	Happy	Sad	Surprised	Fearful	Disgusted	Angry
Real	76.4	93.4	57.8	46.0	46.4	30.4	57.2
StarGAN	92.6	96.1	96.9	95.4	94.7	95.1	95.8
GANmut	87.8	99.2	96.5	91.6	87.5	77.4	97.9
GGANmut	95.8	99.6	99.9	99.1	97.6	97.8	99.5
GANimation	78.4	96.4	76.0	76.0	59.7	52.4	79.6
SMIT	87.4	98.7	87.5	85.5	80.0	78.8	84.6

Table 1: Classification accuracy of generated images with different method. **Real** is the classification error of real images, using as ground truth the test set annotation.

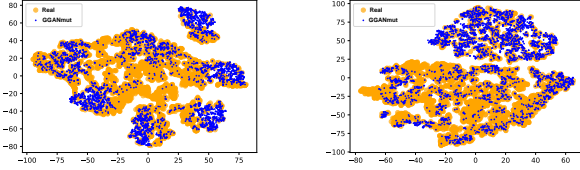


Figure 1: t-SNE plots of VGG features extracted from images generated by GGANmut (blue) and from real images (orange). Left: Conditioning around models of mean μ_c . Right: conditioning uniformly selected samples from Z .

ten less intense and recognisable, reason for which we think the FED score is higher than other methods. On the other hand, if we condition sampling from the entire Z , complex emotions, mixing up 4 basic emotion, will be generated. These expressions, albeit realistic, are even less common, further increasing the FED score (see right plot of Figure 1). Hence, what penalizes GGANmut in the FED test, it is, actually, its strength. In other words, its ability to reproduce the desired emotion as shown from ERE score (Table 2 of the main paper) and classification error (Table 1).

6. More qualitative results

We present more qualitative results for the Gaussian model. In Figure 2 and Figure 3, we created expressions that are difficult to classify inside only one categorical emotion. In Figure 4 we show how the model manages to create different emotion nuances by slight expression modification. And in Figure 5 we show that our model, (taking different conditional codes z_1 and z_2) can produce different expressions even if the label distributions p_{z_1} and p_{z_2} are practically the same. More qualitative results of our two models, in comparison with the state-of-the-art methods, are shown in Figure 6, 7, 8. In Figure 9, 10, 11, we provide more samples synthesized by our linear model. As it can be noticed, with a lower ρ emotions are vague, and expression change smoothly. With $\rho = 1$ emotions become more intense, and expression may change sharply.

7. StarGAN v2

We attempted to produce multiple expressions for the same categorical label using StarGAN-v2. However, as shown in Figures 12 and 13, the method fails to leverage random styles to produce different expressions. This is the reason why we adopted SMIT as our baseline. Note that the StarGAN-v2 was not originally designed for emotion manipulation. Further discussion in this regard can be found in the related works of our paper.

8. Discussion

We observed that the ambiguous supervision of challenging AffectNet dataset – that involves plethora of spontaneous emotions – leads the existing methods to perform poorly. Furthermore, attention-based methods are observed to create artifacts for non-frontal and occluded faces. This particularly turns out to be the case when GANimation, which was originally designed for AU-based control, is trained for categorical emotions. The quality may have been further degraded due to the unreliable annotations of AffectNet. A similar argument can also be made in regard to SMIT. After all, both the original designs aim at somewhat a different goal (please refer to the related works of our paper). The user study is avoided as we observe that the proposed method is clearly better than the competitors in almost all the cases. Please, refer to more qualitative results provided in the supplementary materials. On other small scale datasets, when we tested our method, although a similar behaviour was observed, the achieved improvements were not very prominent. We believe, this is due to the fact that the existing small scale datasets are either non-representative of the spontaneous emotions or are not large enough to model the complex gamut of human emotions. Note that, when the proposed method is restricted, by avoiding intensity and fusion search, it becomes very similar to StarGAN. Therefore, our method can always perform at least as good as the StarGAN. Our desire to model spontaneous emotions and associated issues with the labels makes the AffectNet our ideal candidate. We refrain to report the results of our method on small and controlled (or

curated by avoiding the ambiguous emotions) datasets, as we think that may distract readers from our original intent of modeling spontaneous emotions. Moreover, the presented results are obtained for the most challenging (from the data availability point of view) setup.

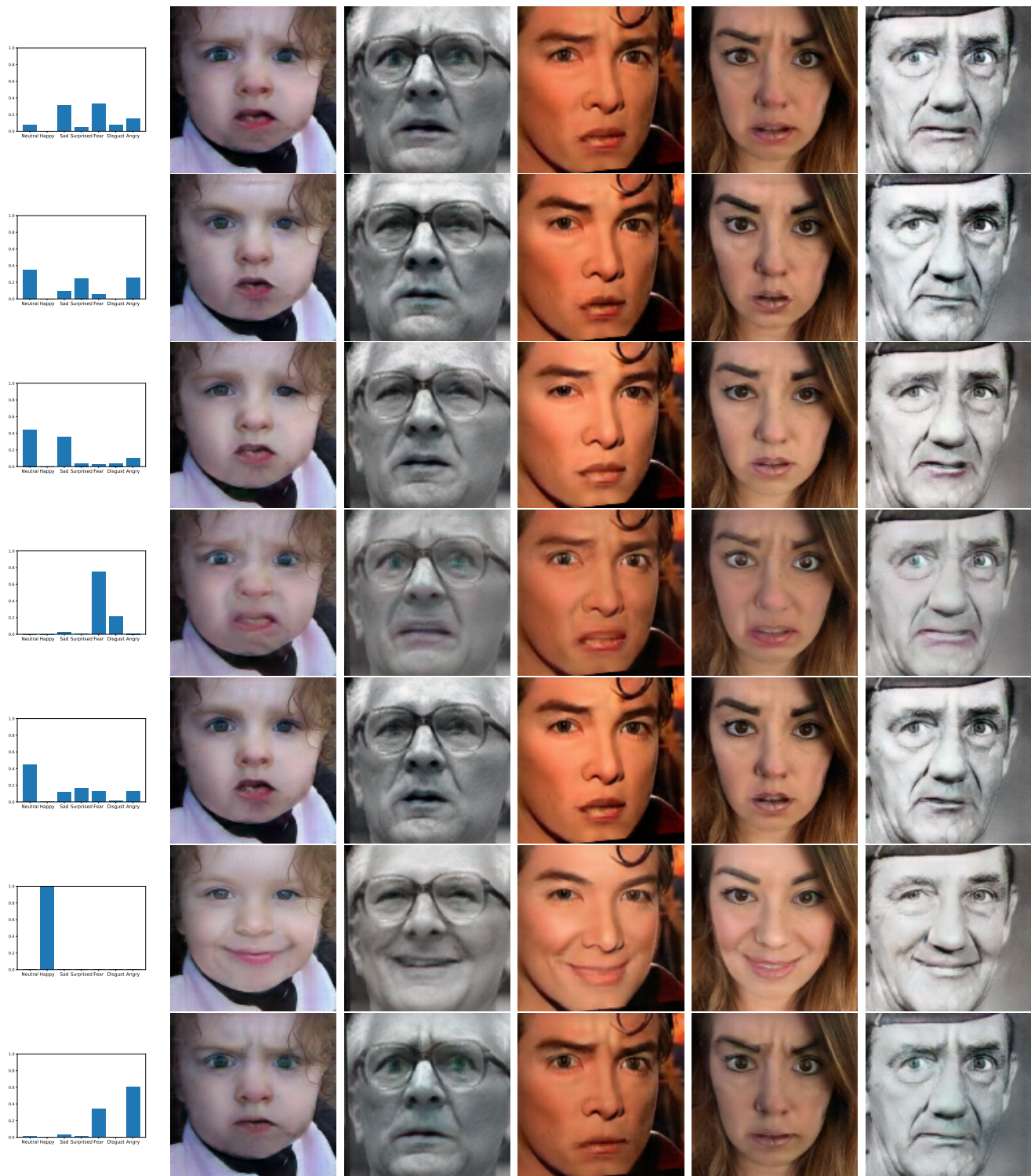


Figure 2: Expression generated using the Gaussian model. In the first column it is provided the label distribution associated with the used conditional code. These images are representative to the complex emotions, which may not be easily classified to basic categorical emotions.

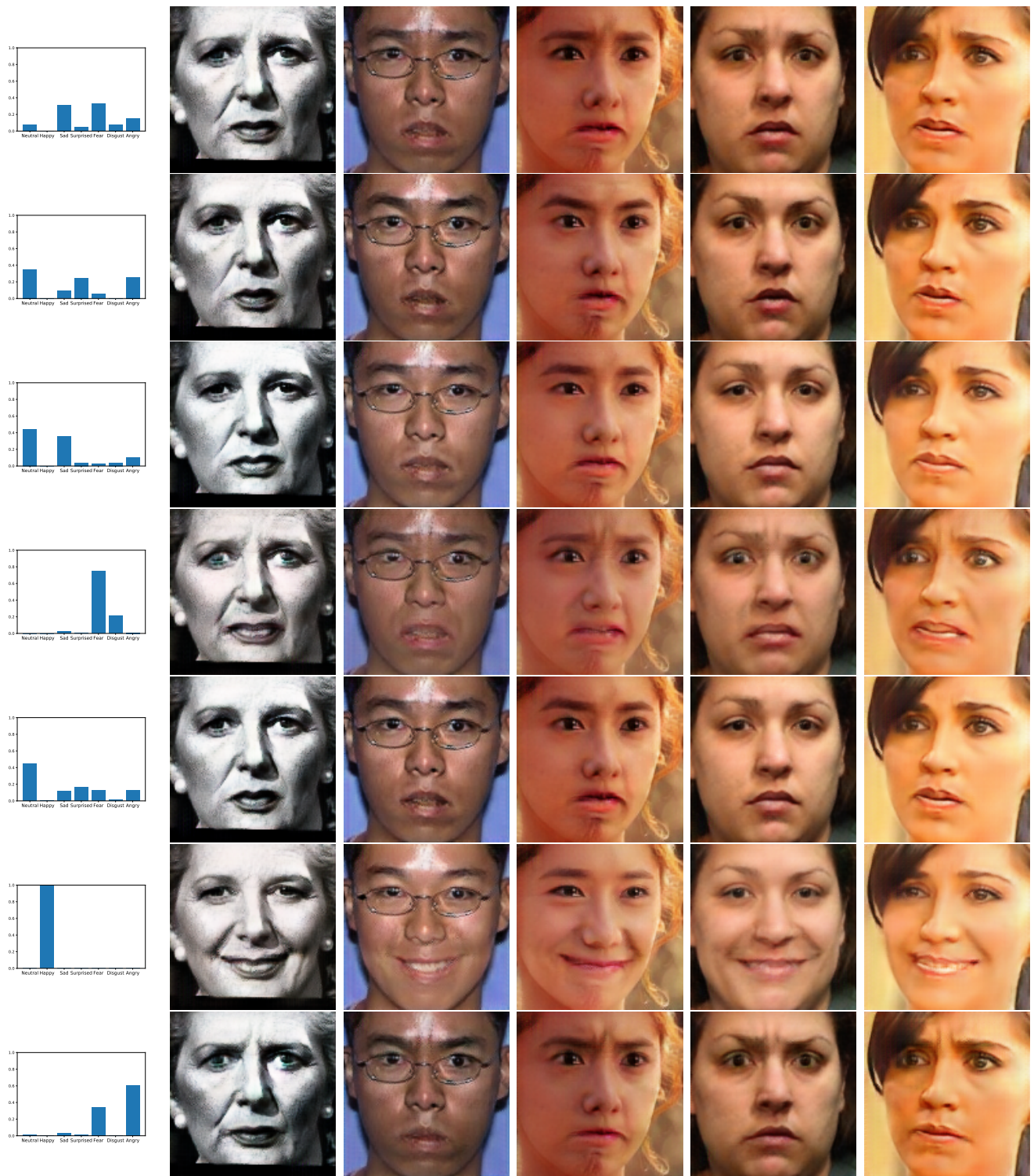


Figure 3: Expression generated using the Gaussian model. In the first column it is provided the label distribution associated with the used conditional code. These images are representative to the complex emotions, which may not be easily classified to basic categorical emotions.

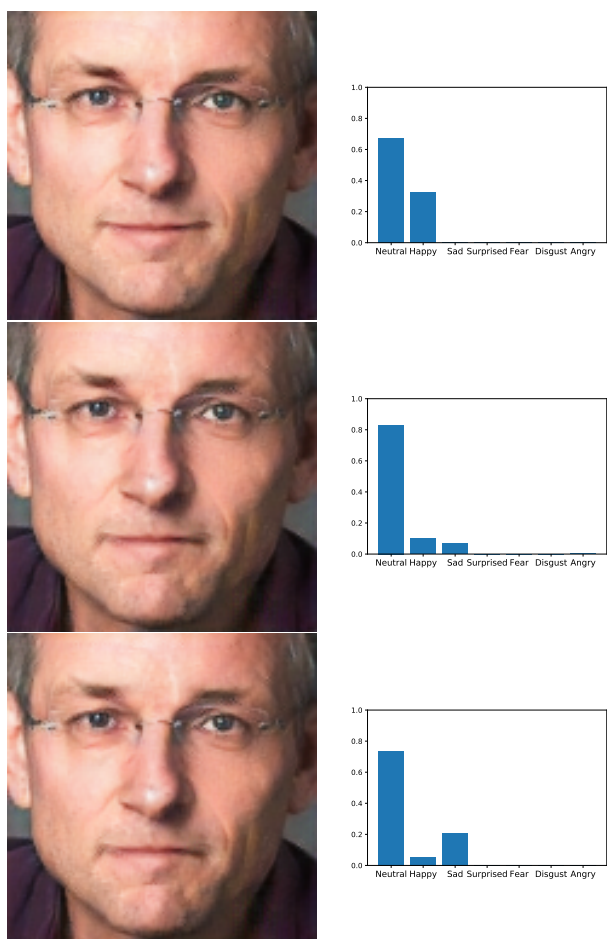


Figure 4: Expressions generated with the Gaussian model. Even though the expression are pretty similar, they convey different emotional nuances.

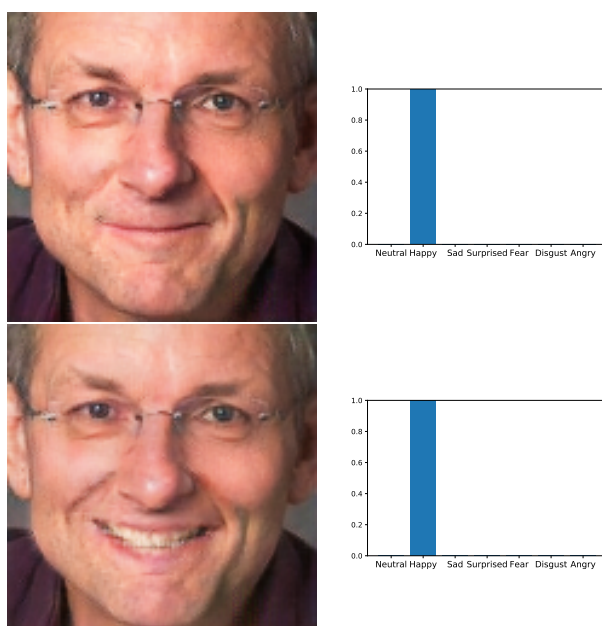


Figure 5: Expressions generated with the Gaussian model. Different conditional codes, associated with the same label distribution, can produce different expressions.

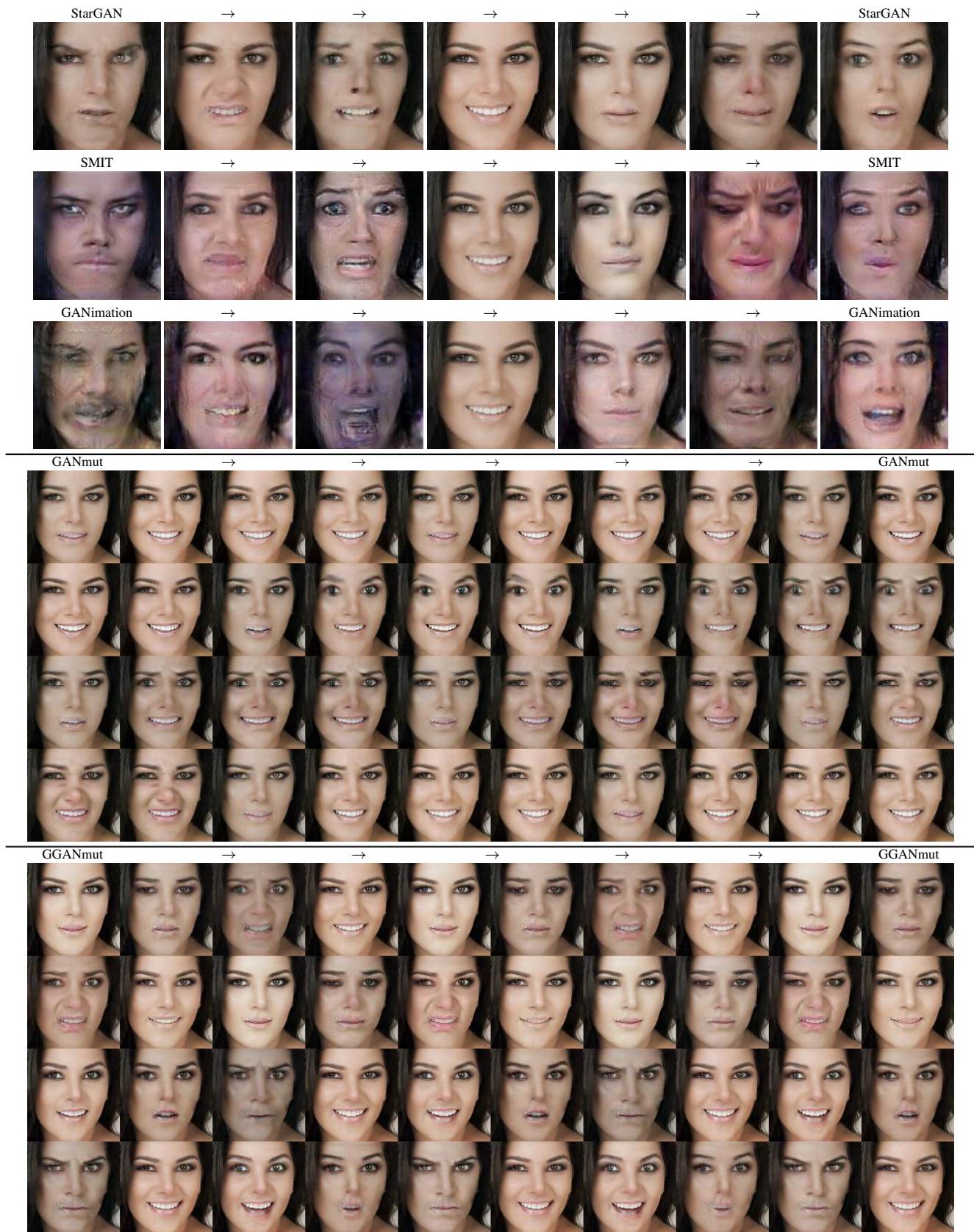


Figure 6: Images synthesized by different methods. Top: basic categorical emotions synthesized by StarGAN, SMIT, and GANimation. Randomize emotions form emotion gamut using our methods: GANmut (middle) and GGANmut (bottom).

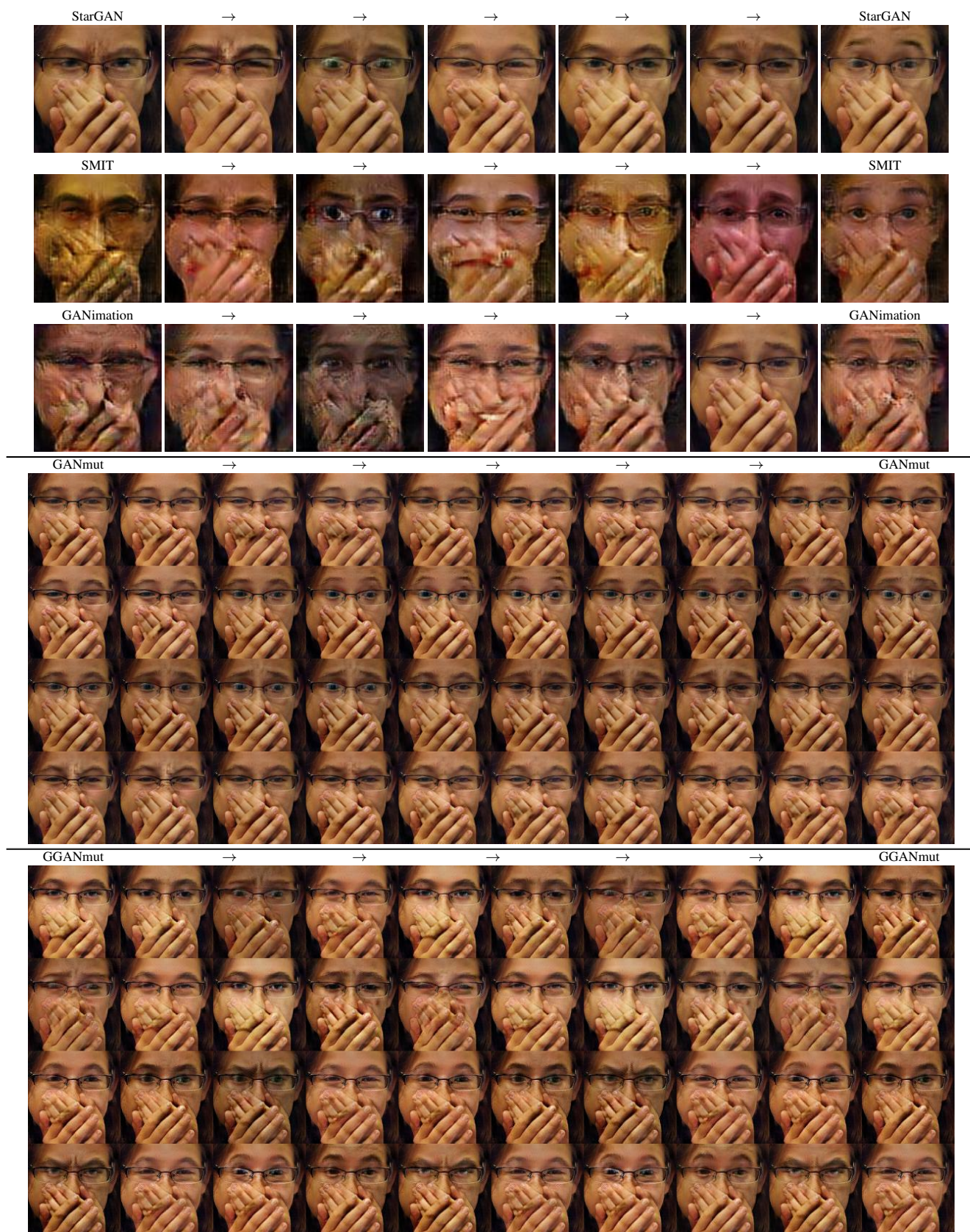


Figure 7: Images synthesized by different methods. Top: basic categorical emotions synthesized by StarGAN, SMIT, and GANimation. Randomize emotions form emotion gamut using our methods: GANmut (middle) and GGANmut (bottom).

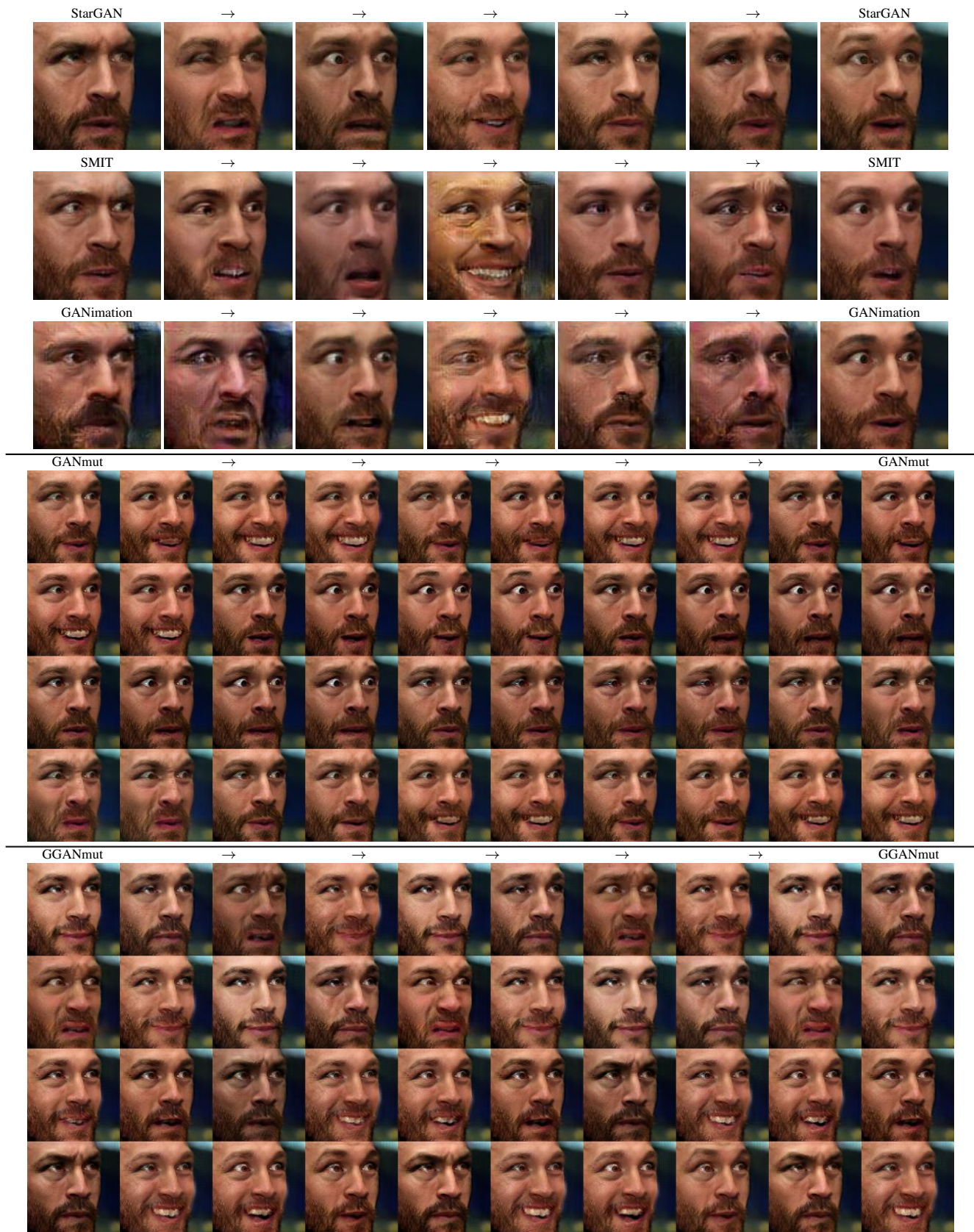


Figure 8: Images synthesized by different methods. Top: basic categorical emotions synthesized by StarGAN, SMIT, and GANimation. Randomize emotions form emotion gamut using our methods: GANmut (middle) and GGANmut (bottom).

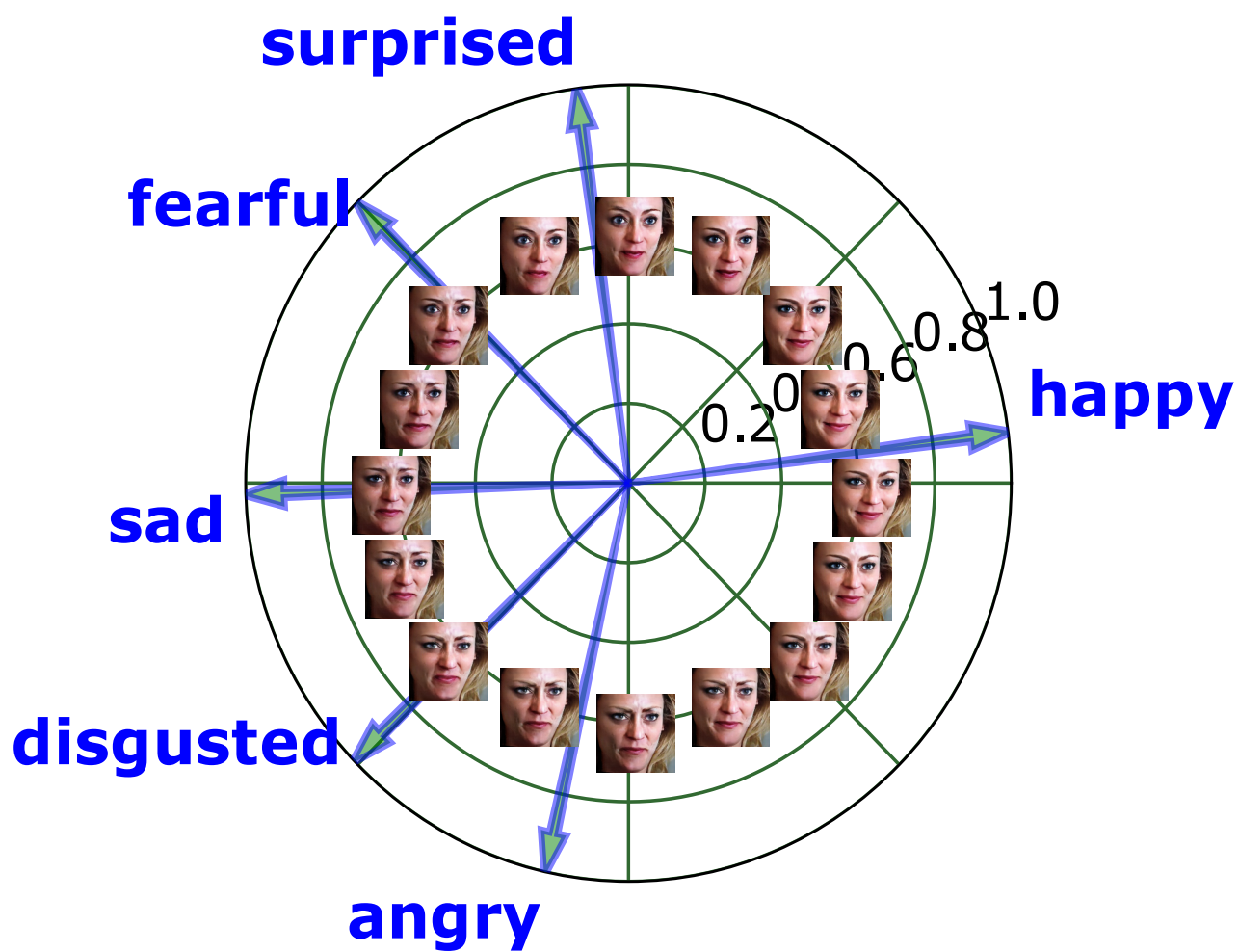


Figure 9: Wheel of emotion generated by the GANmut. It was set $\rho_1 = 0.6$.

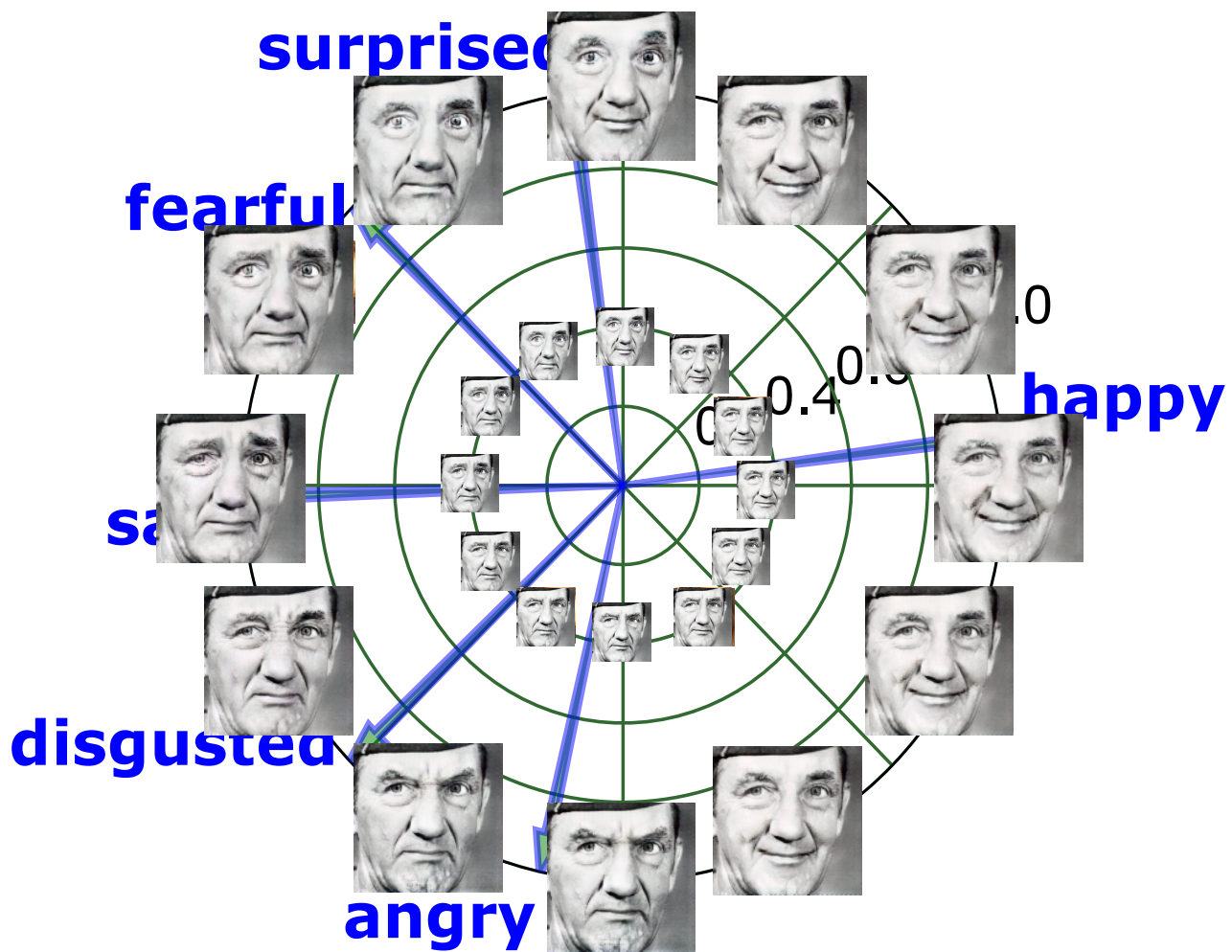


Figure 10: Wheel of emotion generated by the GANmut. It was set $\rho_1 = 0.4$, and $\rho_2 = 1$.

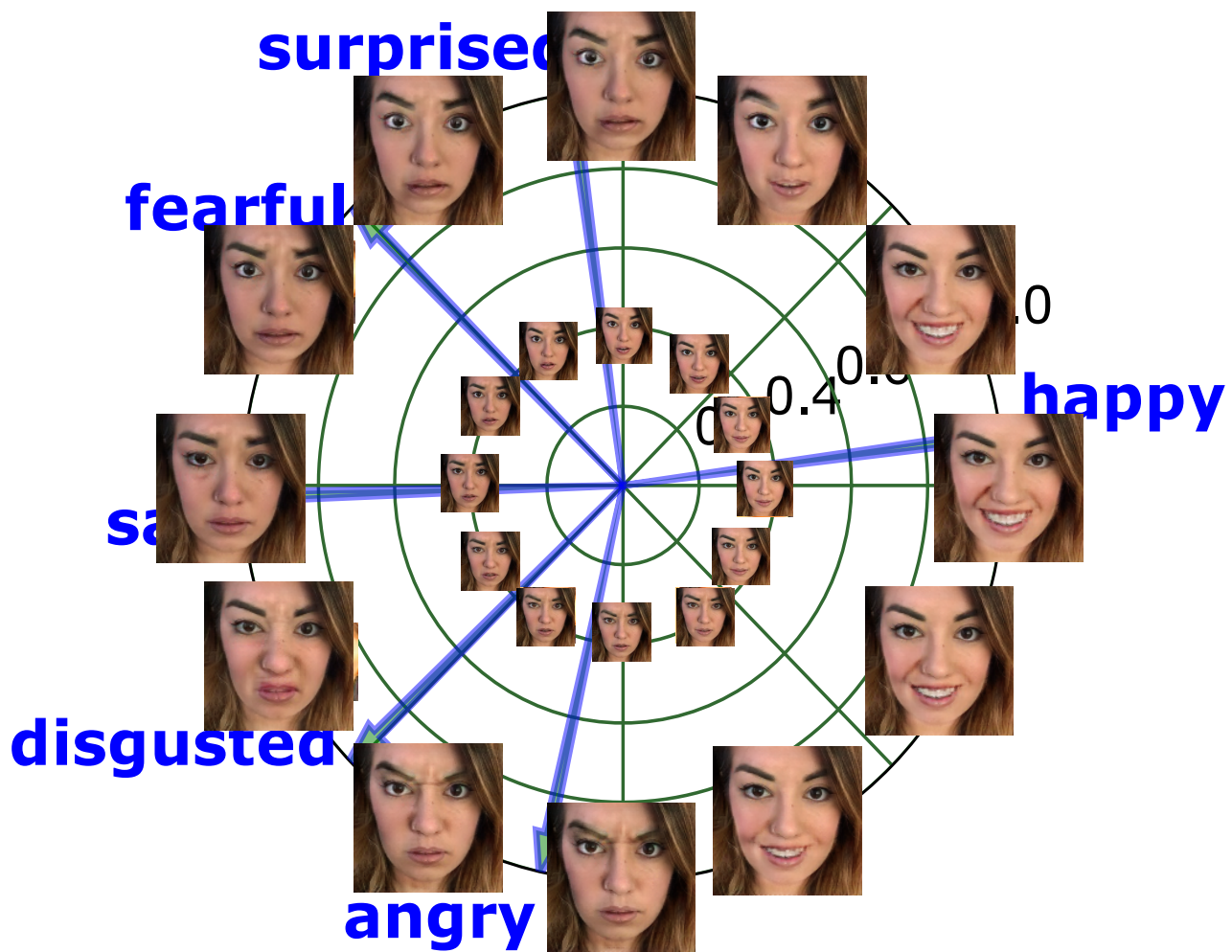


Figure 11: Wheel of emotion generated by the GANmut. It was set $\rho_1 = 0.4$, and $\rho_2 = 1$

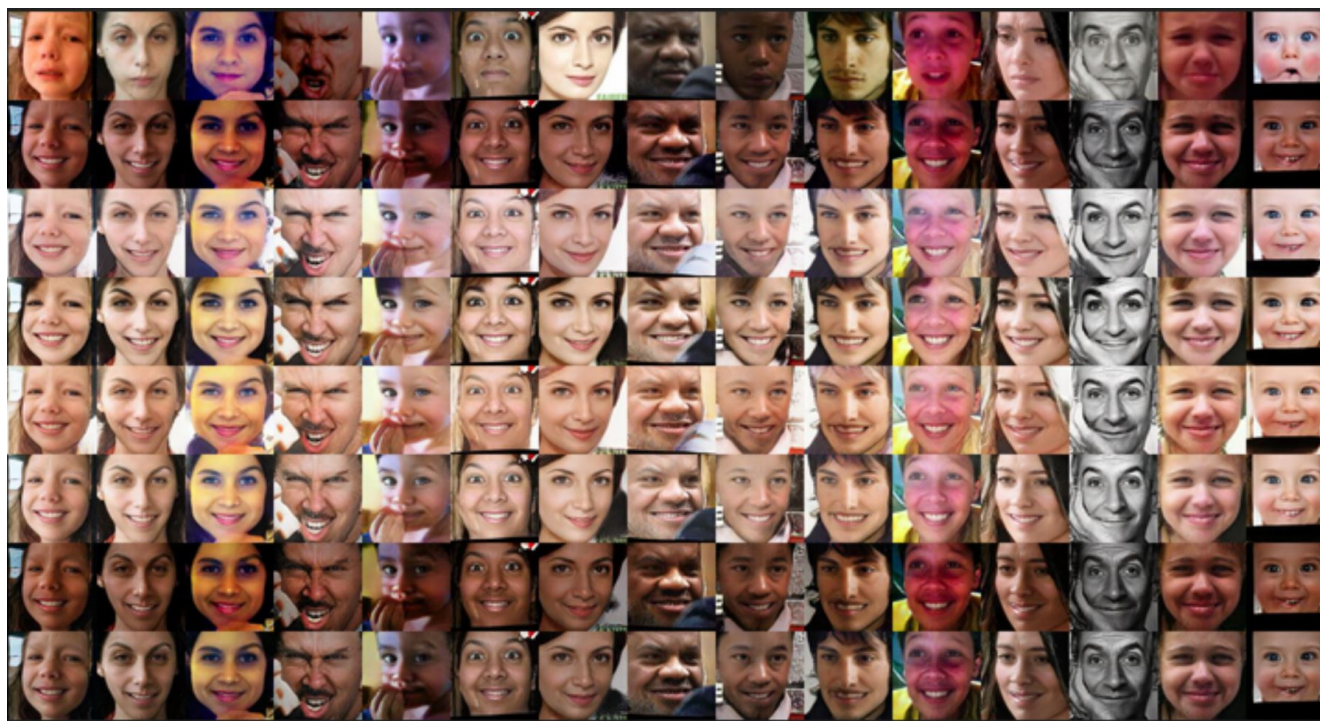


Figure 12: Happy expressions generated with StarGAN-v2. The first row is the input. Different rows correspond to different random styles.

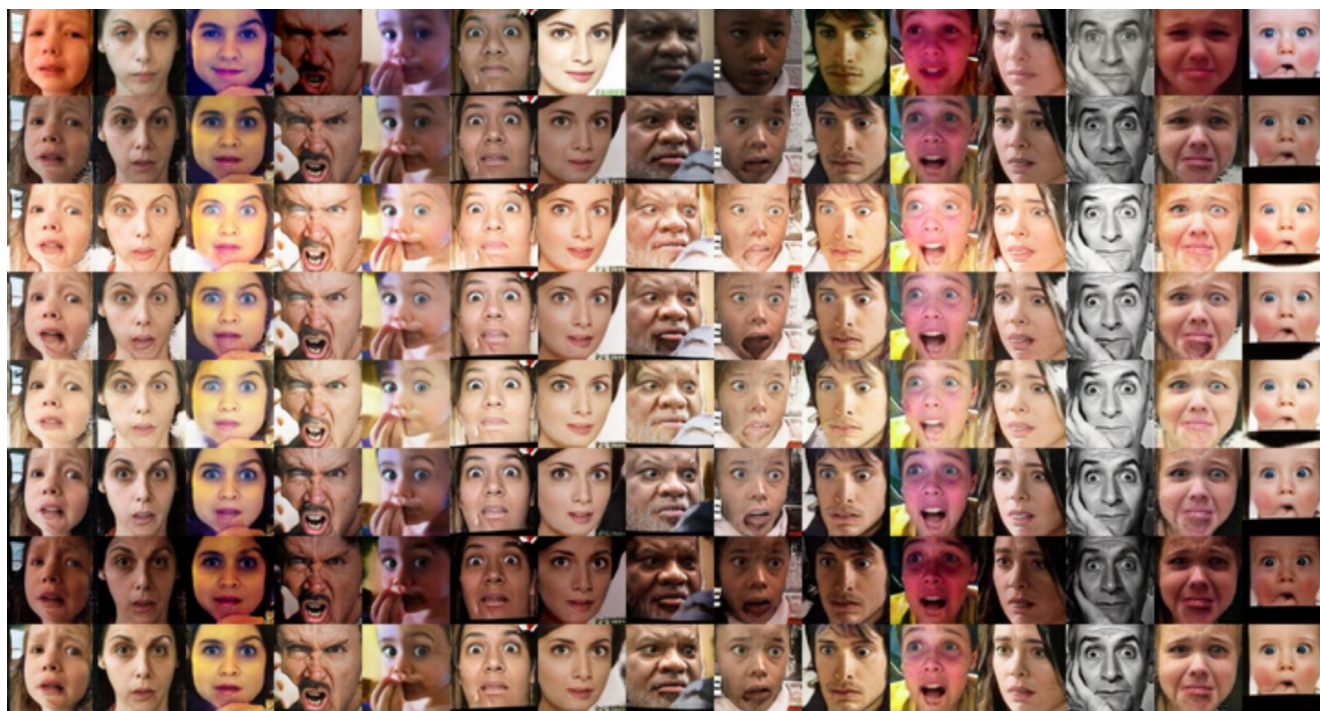


Figure 13: Fearful expressions generated with StarGAN-v2. The first row is the input. Different rows correspond to different random styles.