A Unified and Interpretable Emotion Representation and Expression Generation (Supplementary Materials)

Reni Paskaleva[3*], Mykyta Holubakha[1], Andela Ilic[2], Saman Motamed[1], Luc Van Gool[1,2], Danda Paudel[1]
[1]INSAIT, Sofia University, Bulgaria [2]ETH Zurich, Switzerland
[3]First Private Mathematical High School, Sofia, Bulgaria
firstname.lastname@insait.ai, anilic@student.ethz.ch

This document has of three major sections consisting of additional, (i) details about the user study; (ii) implementation details; and (iii) qualitative results;, in the respective sections. Before proceeding the next, we first provide different view-point visualizations of the proposed 3D emotion model in Figure 1 and 3.

1. User (psychologist) Study

We performed two studies, each of 2D AV and proposed 3D C2A2 models, both consisting of 315 images each. These 315 images were divided into five parts, each containing 63 images. A user could choose any image from any part, one by one, and rate that image in the scale between 1 to 5, with higher being better quality. We ensured that the users vote only once per image, by providing the appropriate interfaces. For creating the 3D interface we used Three.js and for the 2D one we used javascript + html with no external libraries. For the backend of both we used FastAPI. Two screenshots of the designed interfaces are shown in Figure 2. The 3D A2C2 model could be rotated for better visibility. The further details are available at:
https://emotion-diffusion.github.io/

As mentioned, eight expert psychologists were asked to annotate all 2x315 images. Once the ratings were obtained, we obtain both overall and class-wise scores for the compound emotions. For class-wise scores, we assign the samples based on their nearest compound emotion classes.

2. Implementation Details

2.1. DreamBooth and GANmut framework details

For our model we finetune Stable Diffusion 1.4. We set the regularization weight of the model to 1.0 and the learning rate to 1.0e-6. The sampling steps are fixed to 100. We use the beta scheduler with parameters linear start=0.00085 and linear end=0.0120. Please, refer to the original implementation of DreamBooth at https://github.com/XavierXiao/Dreambooth-Stable-Diffusion for further details. Similarly, we follow the original implementation of GANmut at https://github.com/stefanodapolito/GANmut, for the hyper-parameters and the backbone selection. The details of the loss modification is given in the following subsection.

2.2. During Z recovery

As mentioned, our method of learning Z relies on the method developed in GANmut [1]. In the notations below, z refers to the conditional code, while c' and c denote

Emotion	2D-AV	3D (Ours)
Happy	4.28	3.90
Sad	3.76	3.76
Surprised	4.10	4.09
Fearful	3.47	3.98
Disgusted	3.57	4.50
Angry	3.42	4.49
Neutral	3.91	3.66
Happily-Sad	3.12	4.18
Happily-Surprised	4.34	4.36
Happily-Fearful	3.41	4.32
Happily-Disgusted	2.71	3.51
Sadly-Fearful	3.00	4.06
Sadly-Angry	2.67	4.18
Sadly-Surprised	2.50	4.02
Sadly-Disgusted	3.98	4.57
Fearfully-Angry	4.15	4.12
Fearfully-Surprised	3.97	4.52
Fearfully-Disgusted	3.09	4.52
Angrily-Surprised	3.67	4.74
Angrily-Disgusted	2.17	4.73

Table 1. Scores obtained for different compound emotion classes from our user study. As expected, 2D-AV models perform well only on few basic emotions. Our 3D C2A2 model performs significantly better overall on the preference of expert psychologists.

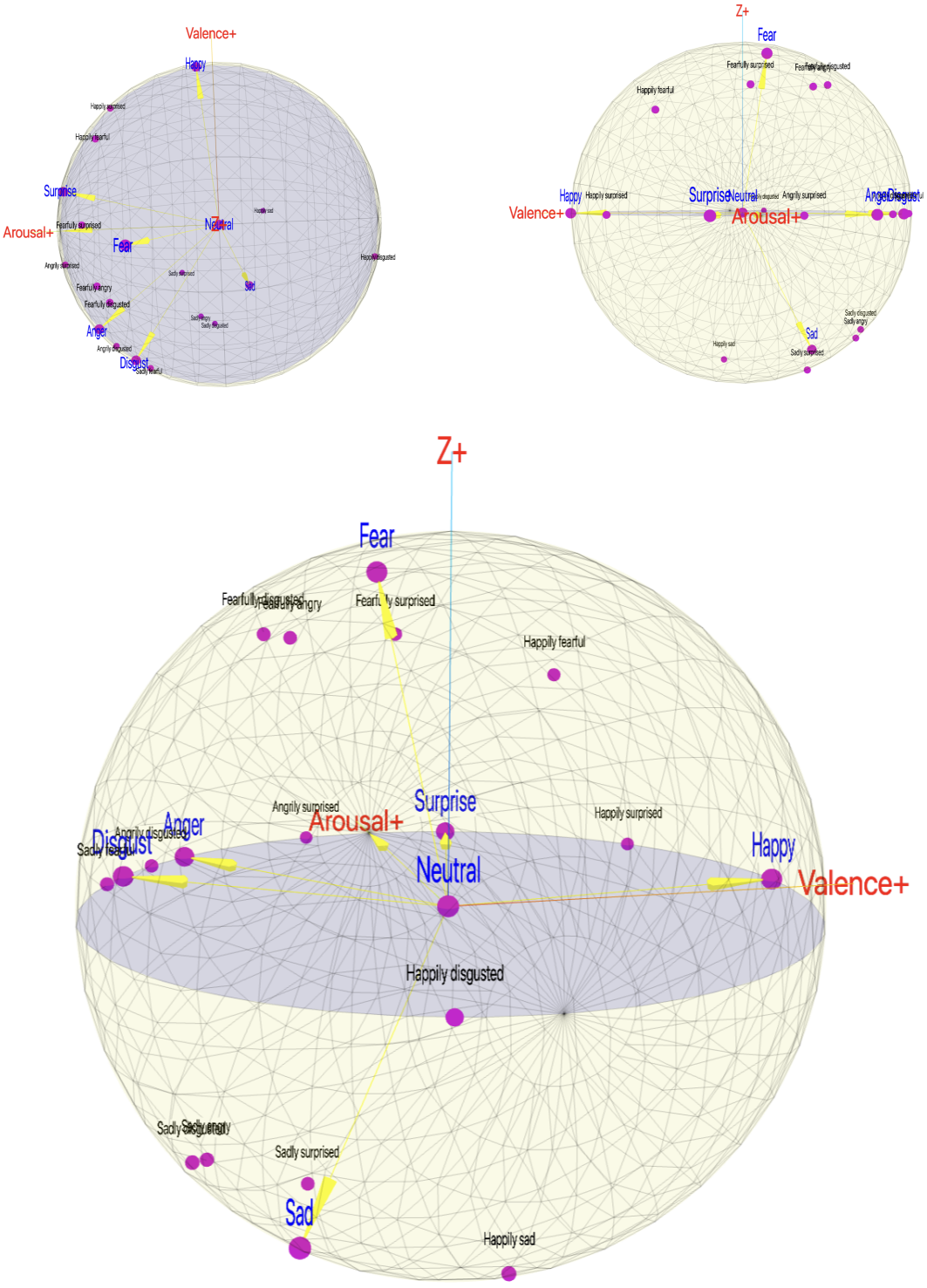

Figure 1. The top (top-left), front (top-right), and best (bottom) views of the proposed 3D C2A2 model. Note that the Fear and Sad are lifted to 3D from the 2D AV model. This allows a more complete coverage of the compound emotions without losing interpretability.

the original and target labels, respectively. D_{src}, D_{cls} and D_{coor} are the outputs of the discriminator network. $D_{src}(x)$ represents how likely the input image is real. $D_{cls}(c|x)$ shows the probability of an input image x to

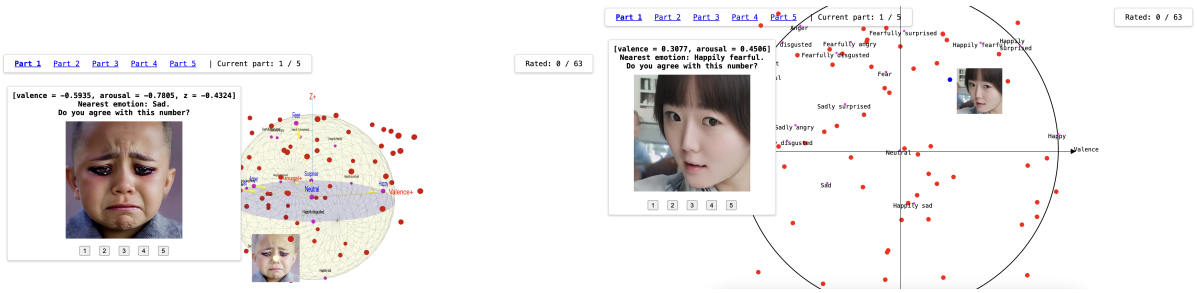

Figure 2. Screenshots of the user interface (Left: 3D C2A2 and right: 2D AV) used for our study. Interactive links are provided above.

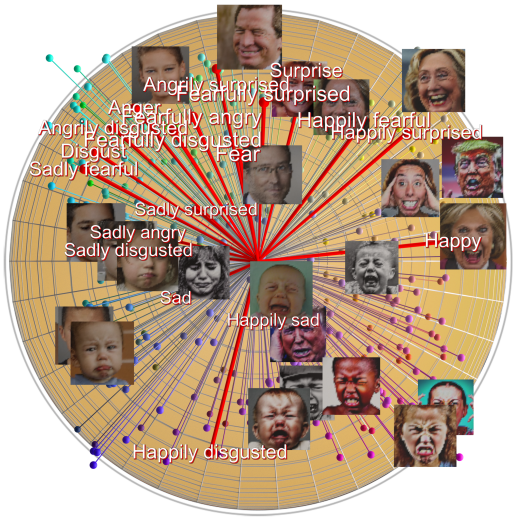

Figure 3. An Example image collection in the 3D C2A2 space.

be classified into each of the seven emotion categories, while $\hat{\rho}$ denotes the estimated expression strength based on the class to which image x belongs most likely. D_{coor} estimates the coordinates of the image x in the VA plane. $\hat{x}$ is used for Wasserstein loss and denotes the image sampled along the line that connects a real and a generated image. Now, we are ready to list the used losses.

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

$$\mathcal{L}^f_{cls} = \mathbb{E}_{x,c,\rho}[-\log D_{cls}(c|G(x, z_{c,\rho}(\theta_c)))]. \quad (2)$$

$$\mathcal{L}^r_{cls} = \mathbb{E}_{x,c'}[-\log D_{cls}(c'|x)]. \quad (3)$$

$$\mathcal{L}_{info} = \mathbb{E}_{x,z}\left[||\, D_{coor}(G(x,z))) - z||_2^2\right]. \quad (4)$$

$$\mathcal{L}_{\rho} = \mathbb{E}_{x,z}\left[||\, \hat{\rho}(G(x, z_{.,\rho}))) - \rho||_2^2 \mathbb{1}_{\rho>0.2}\right]. \quad (5)$$

$$\mathcal{L}_{rec} = \mathbb{E}_{x,z}\left[||\, x - G(G(x,z), D_{coor}(x))||_1\right]. \quad (6)$$

Unlike GANmut, our losses are formulated in 3D space to capture much more compound and complex emotions. The full objective functions of the generator and discriminator of C2A2 model are:

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^r_{cls} + \lambda_{info_D}\mathcal{L}_{info} + L_{av} + L_{AU_Y}. \quad (7)$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^f_{cls} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{info_G}\mathcal{L}_{info} + \lambda_{\rho}\mathcal{L}_{\rho}. \quad (8)$$

The above loses for discriminator and generator are alternatively applied, and the complete reconstruction loss is computed only for the fake images. For real images, the Z axis gets supervised using the action units mapping as discussed in the method section. Please, refer the main paper regarding the final recovery of the pseudo-labels along Z.

3. Qualitative Results

In this section, we provide various qualitative results. These qualitative results are presented in Figure 4 - 45. Although the descriptions of the figures are provided in their respective captions, here we provide a summary of them in Table 2. Overall, the qualitative results also support the benefits of proposed 3D C2A2 emotion model.

References

[1] Stefano d'Apolito, Danda Pani Paudel, Zhiwu Huang, Andres Romero, Luc Van Gool. Ganmut: Learning interpretable conditional space for gamut of emotions. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 568–577, 2021. 1

Description	Model	Figure
Scenes	3D C2A2	Fig.4
Attributes	3D C2A2	Fig.5
Emotion axis	3D C2A2	Fig. 6-13
Emotion axis	2D AV	Fig. 14-21
Circle Walk	3D C2A2	Fig. 22-25
Circle Walk	2D AV	Fig. 26-29
Interpolation	3D C2A2	Fig. 30-37
Interpolation	2D AV	Fig. 38-45

Table 2. Summary of the qualitative results.

Figure 4. Expressions generated by our method with different scenes in the background.

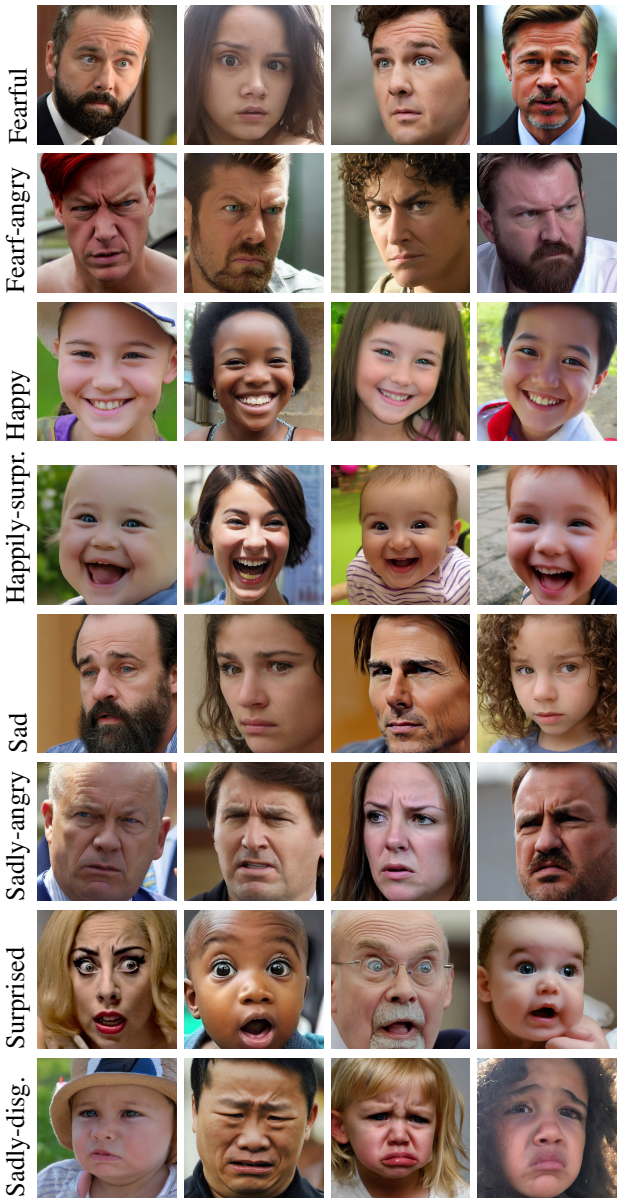

Figure 5. Thanks to our method we can generate meaningful emotions with various different attributes - different hair and eye colour, accessories, age, ethnicity. We can even generate celebrities with various expressions.

Figure 6. 3D model - interpolation along the axis of happy emotion

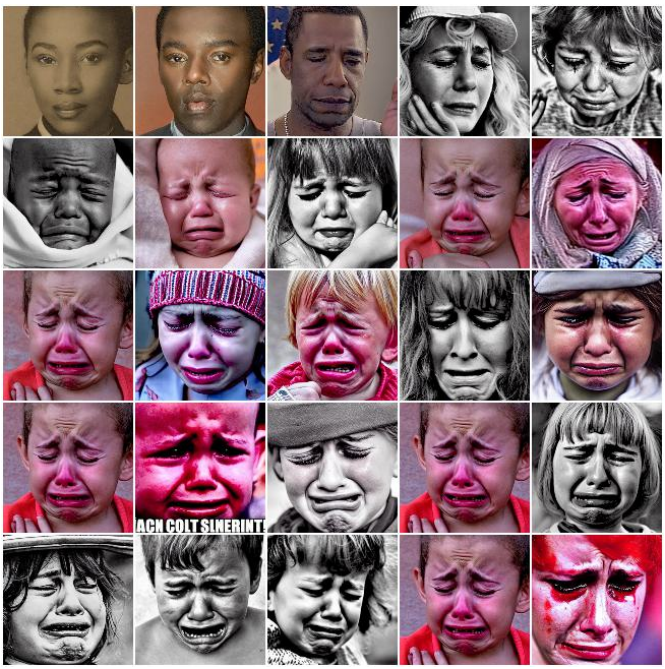

Figure 8. 3D model - interpolation along the axis of sad emotion

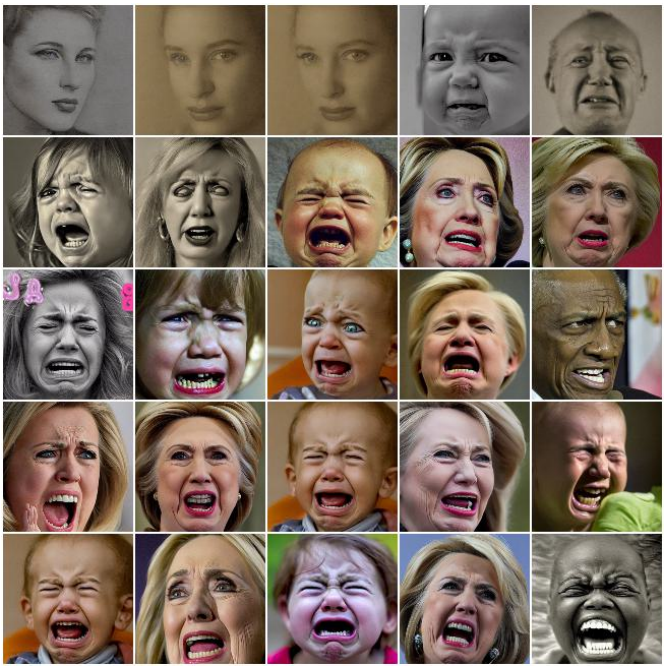

Figure 7. 3D model - interpolation along the axis of fearful emotion

Figure 9. 3D model - interpolation along the axis of surprised emotion

Figure 10. 3D model - interpolation along the axis of fearfully-disgusted emotion

Figure 12. 3D model - interpolation along the axis of sadly-angry emotion

Figure 11. 3D model - interpolation along the axis of happily-fearful emotion

Figure 13. 3D model - interpolation along the axis of angrily-surprised emotion

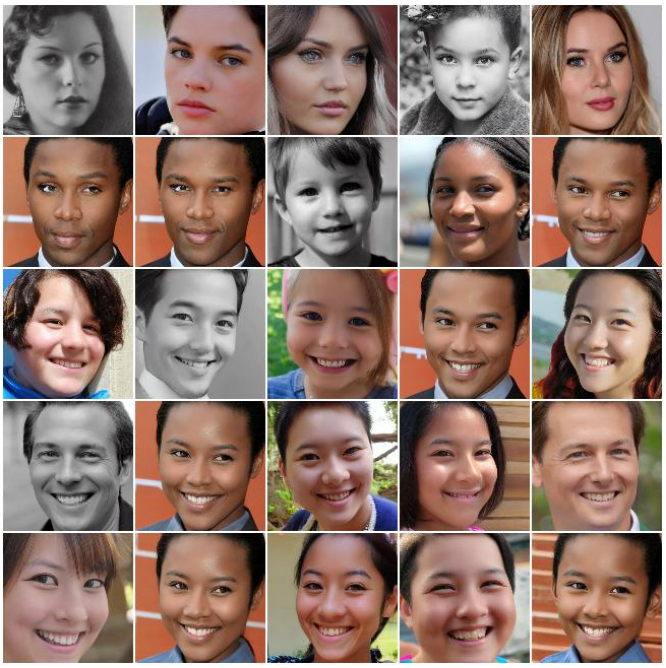

Figure 14. 2D model - interpolation along the axis of happy emotion

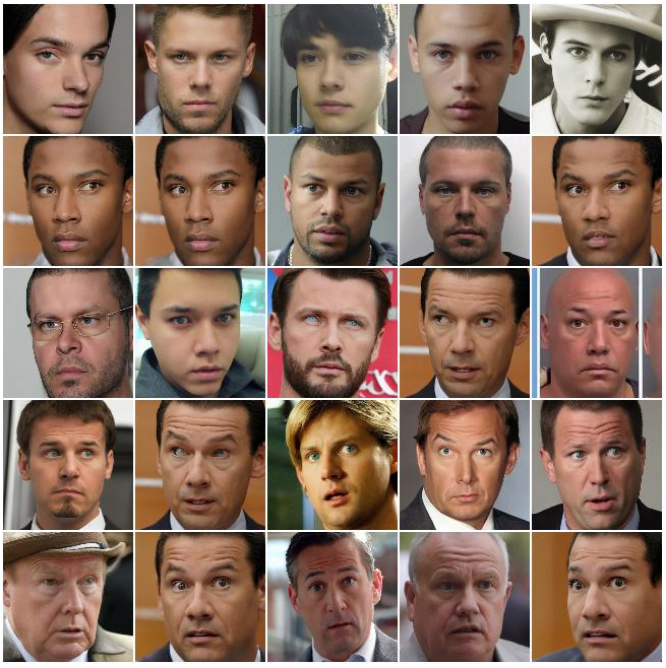

Figure 15. 2D model - interpolation along the axis of fearful emotion

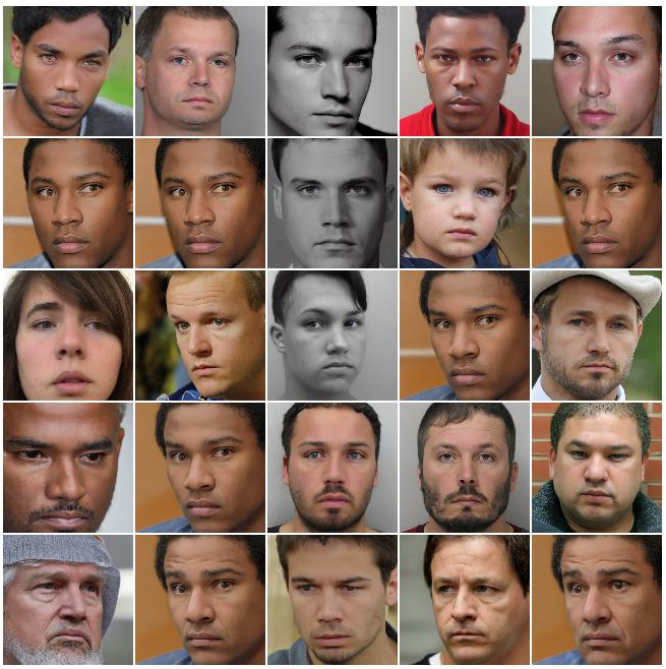

Figure 16. 2D model - interpolation along the axis of sad emotion

Figure 17. 2D model - interpolation along the axis of surprised emotion

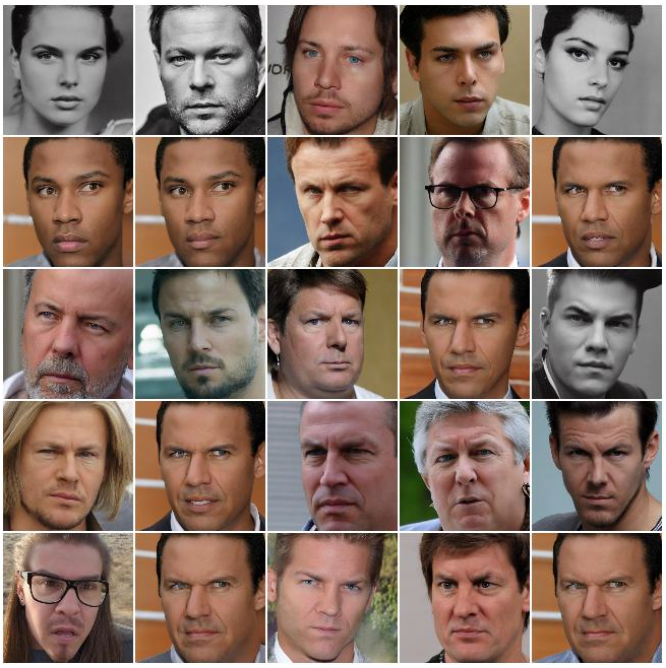

Figure 18. 2D model - interpolation along the axis of fearfully-disgusted emotion

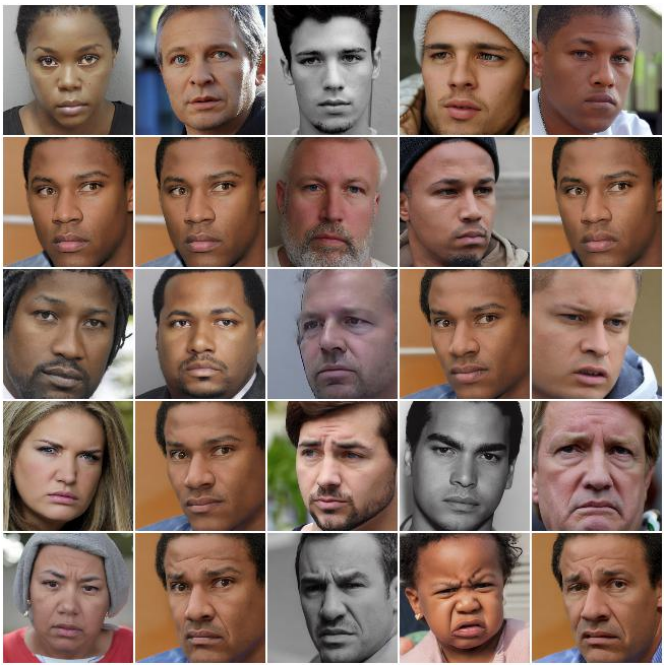

Figure 20. 2D model - interpolation along the axis of sadly-angry emotion

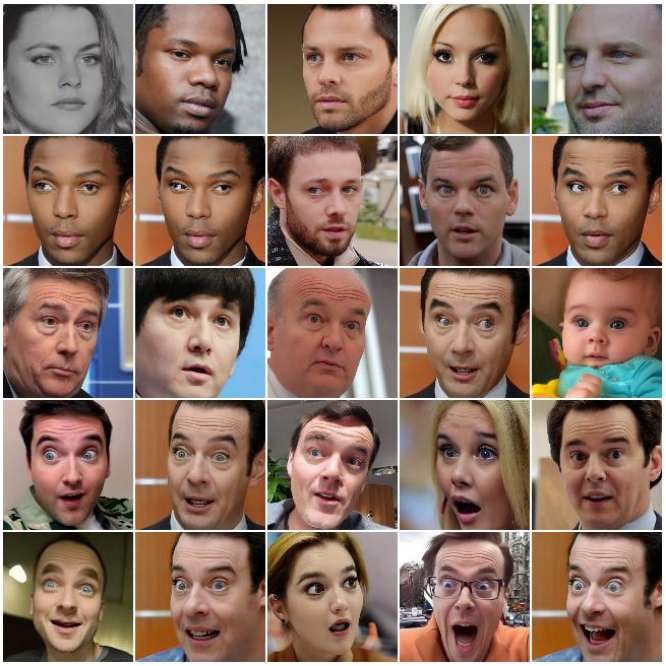

Figure 19. 2D model - interpolation along the axis of happily-fearful emotion

Figure 21. 2D model - interpolation along the axis of angrily-surprised emotion

Figure 22. 3D model - Circle θ=0 ϕ=0 r=0.8

Figure 24. 3D model - Circle θ=0.4901 ϕ=0.9801 r=0.6

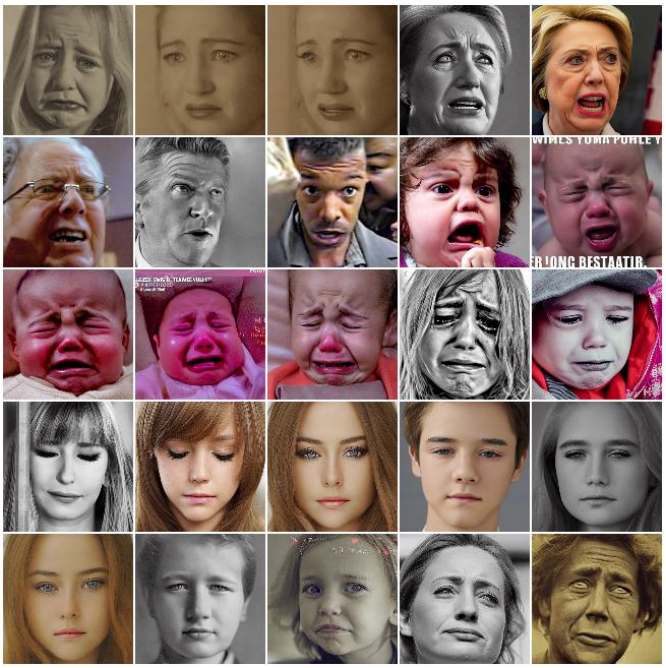

Figure 23. 3D model - Circle θ=2.3 ϕ=3.761 r=0.2

Figure 25. 3D model - Circle θ=1.177 ϕ=5.974 r=1.0

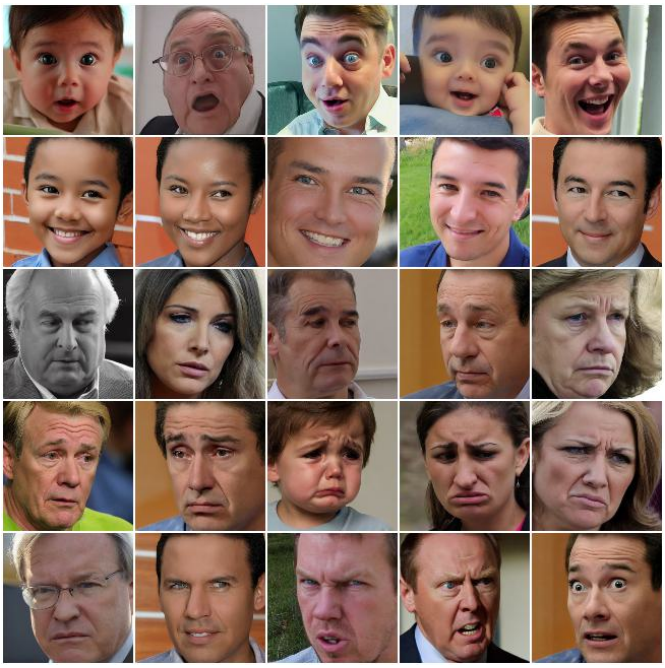

Figure 26. 2D model - Circle θ=0 ϕ=0 r=0.8

Figure 28. 2D model - Circle θ=0.4901 ϕ=0.9801 r=0.6

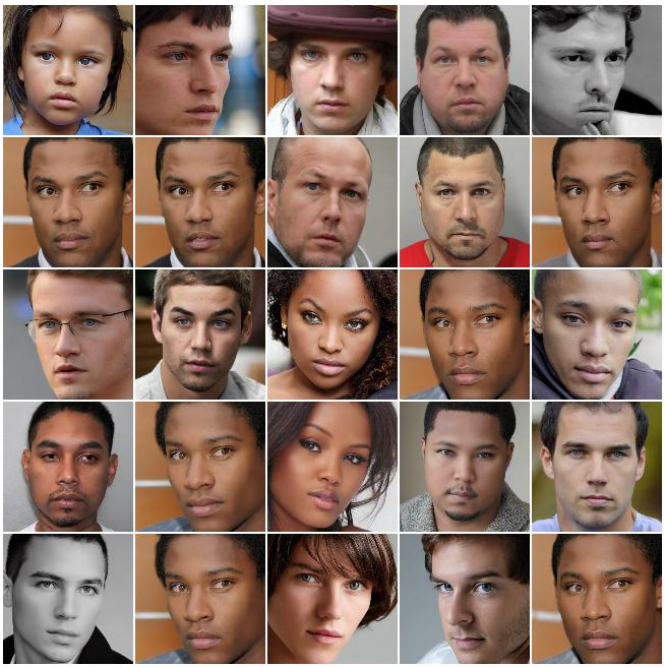

Figure 27. 2D model - Circle θ=2.3 ϕ=3.761 r=0.2

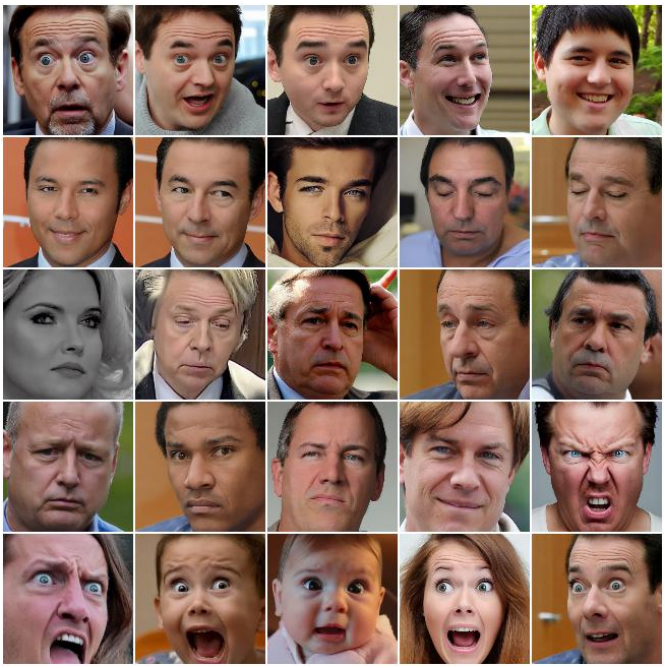

Figure 29. 2D model - Circle θ=1.177 ϕ=5.974 r=1.0

Figure 30. 3D model - interpolation between surprised and angry emotions

Figure 32. 3D model - interpolation between happy and fearful emotions

Figure 31. 3D model - interpolation between fearful and disgusted emotions

Figure 33. 3D model - interpolation between surprised and fearful emotions

Figure 34. 3D model - interpolation between happily-sad and happily-fearful

Figure 36. 3D model - interpolation between happily-fearful and sadly-fearful

Figure 35. 3D model - interpolation between happily-surprised and fearfully-surprised emotions

Figure 37. 3D model - interpolation between happily-surprised and angrily-surprised emotions

Figure 38. 2D model - interpolation between surprised and angry emotions

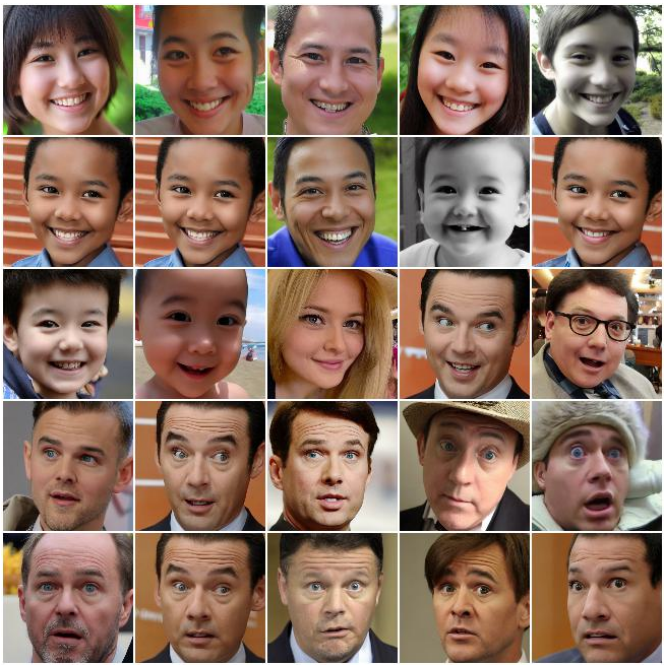

Figure 40. 2D model - interpolation between happy and fearful emotions

Figure 39. 2D model - interpolation between fearful and disgusted emotions

Figure 41. 2D model - interpolation between surprised and fearful emotions

Figure 42. 2D model - interpolation between happily-sad and happily-fearful

Figure 44. 2D model - interpolation between happily-fearful and sadly-fearful

Figure 43. 2D model - interpolation between happily-surprised and fearfully-surprised emotions

Figure 45. 2D model - interpolation between happily-surprised and angrily-surprised emotions