UP-FacE:
User-predictable Fine-grained Face Shape Editing
Supplementary Material

Florian Strohm[*,1,3] Mihai Bâce[*,2] Andreas Bulling[3]
[1]Fraunhofer IPA [2]KU Leuven [3]University of Stuttgart

1. Full Implementation Details

Input to our Transformer encoder is the sequence of vectors $w^+ \in \mathbb{R}^{512\times18}$ concatenated with the 512 dimensional semantic face feature embedding extracted by the embedding layer. The encoder consists of four Transformer encoder layers with four self-attention heads [7] each. The output of the Transformer is input to a linear layer that outputs the semantic manipulation vector s_e. The scaling network takes the current and target face feature values, m_j and m_j^t for a feature j, and the feature's embedding e_j as input. It consists of three linear layers with 32, 32, and one hidden unit(s), respectively, with a ReLU activation function after the first two linear layers and outputs the scaling value k.

To train UP-FacE, we generate data batches by sampling random vectors $z \sim \mathcal{N}(0,1)$ from a standard normal distribution. We generate the corresponding vectors w^+ and images I using the mapping and synthesis network from a pre-trained StyleGAN2 [2] model. For each image, we sample a face feature index j we want to modify from a uniform distribution $j \sim \mathcal{U}(0,23)$. We also sample a desired target feature value m_j^t from a standard normal distribution $m_j^t \sim \mathcal{N}(0,1)$. Furthermore, we use the pre-trained SPIGA landmark detector [4] to extract landmarks and calculate the current face feature values m_j for each generated image I. Since we sample m_j^t from $\mathcal{N}(0,1)$ we normalise m_j to follow a standard normal distribution as well:

$$m_j = m_j - \mu_j/\sigma_j, \tag{1}$$

where μ and σ are the mean and standard deviation for all face features j calculated on the FFHQ dataset, which our StyleGAN2 model was pre-trained on. This ensures that our defined semantic face features are normalised to the same value range when calculating the loss. Using these generated data samples, we run UP-FacE to predict w^+_{edit}. We then use the StyleGAN2 synthesis network to generate I_{edit} and the SPIGA landmark detector to extract the new landmarks and calculate the predicted feature values m_j^p. We calculate the loss as defined in the main paper and propagate the gradients back through SPIGA and StyleGAN to update the parameters of UP-FacE. We optimise UP-FacE for 10^5 steps using the Adam optimiser [3] with a learning rate of 2^{-5} and a batch size of 16. Through empirical testing, we set the weighting scalars of the loss function to $\lambda_{\text{pix}} = 1$, $\lambda_{\text{feat}} = 3$, $\lambda_{\text{SFF}} = 0.005$, $\lambda_{\text{reg}} = 0.1$ and $\lambda_{\text{cor}} = 1$.

[*]Part of this work was done while at University of Stuttgart.

2. Face Feature Correlations

To understand how changing one semantic feature affects others, we compute all 23 feature values for each face in FFHQ [2]. We then estimate pairwise Pearson correlations between all features. Figure 1 shows the full correlation matrix. For visual clarity, we highlight only stronger correlations in the printed values. Most high correlations occur among width-related features, indicating that facial proportions tend to scale jointly.

3. Interactive Face Editing Tool

Inspired by common digital character creation tools [1, 5], we provide a slider-based interface for UP-FacE. Each of the 23 semantic face features is controlled by one slider. The interface is shown in Figure 2. When a slider changes, UP-FacE performs a forward pass and updates the edited face immediately. Because correlated features can change jointly, we re-estimate feature values after each edit and update related sliders. An anonymised video demonstration is available at https://youtu.be/xSXAJP1M3ew.

4. Additional Editing Results

Figure 3 shows additional progressive edits with more examples. Figures 4 to 7 shows additional real-image edits for all 23 semantic features. For each pair, the left image is the original real face inverted into StyleGAN2 latent space [2, 6]. The right image is the edited result produced by UP-FacE. We use strong edit magnitudes to make the direction of each semantic edit clearly visible.

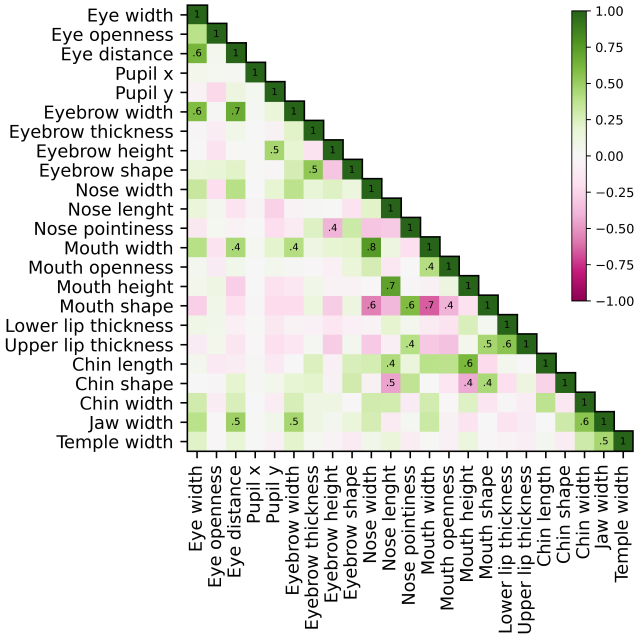

Figure 1. Full Pearson correlation matrix of our 23 semantic face features on FFHQ. Correlations with larger absolute magnitude indicate feature pairs that naturally co-vary.

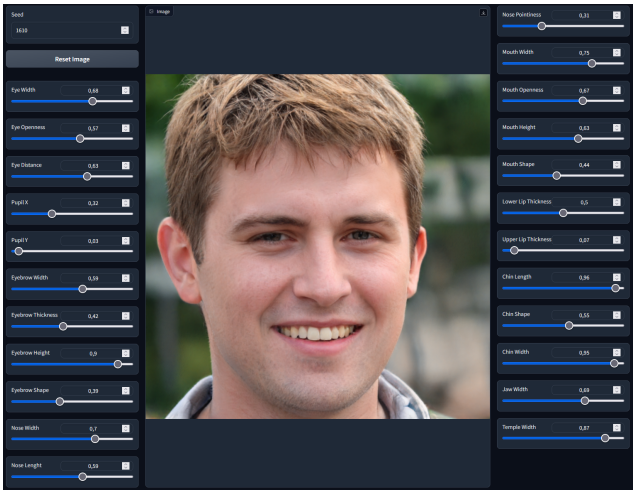

Figure 2. Slider-based user interface for UP-FacE. Users can perform fine-grained semantic edits by directly manipulating feature sliders.

References

[1] Leyde Briceno and Gunther Paul. Makehuman: a review of the modelling framework. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume V: Human Simulation and Virtual Environments, Work With Computing Systems (WWCS), Process Control 20*, pages 224–232. Springer, 2019.

[2] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[4] Andrés Prados-Torreblanca, José M Buenaposada, and Luis Baumela. Shape preserving facial landmarks with graph attention networks. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.

[5] Valentin Schwind, Katrin Wolf, and Niels Henze. Facemaker—a procedural face generator to foster character design research. *Game Dynamics: Best Practices in Procedural and Dynamic Game Content Generation*, pages 95–113, 2017.

[6] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–14, 2021.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

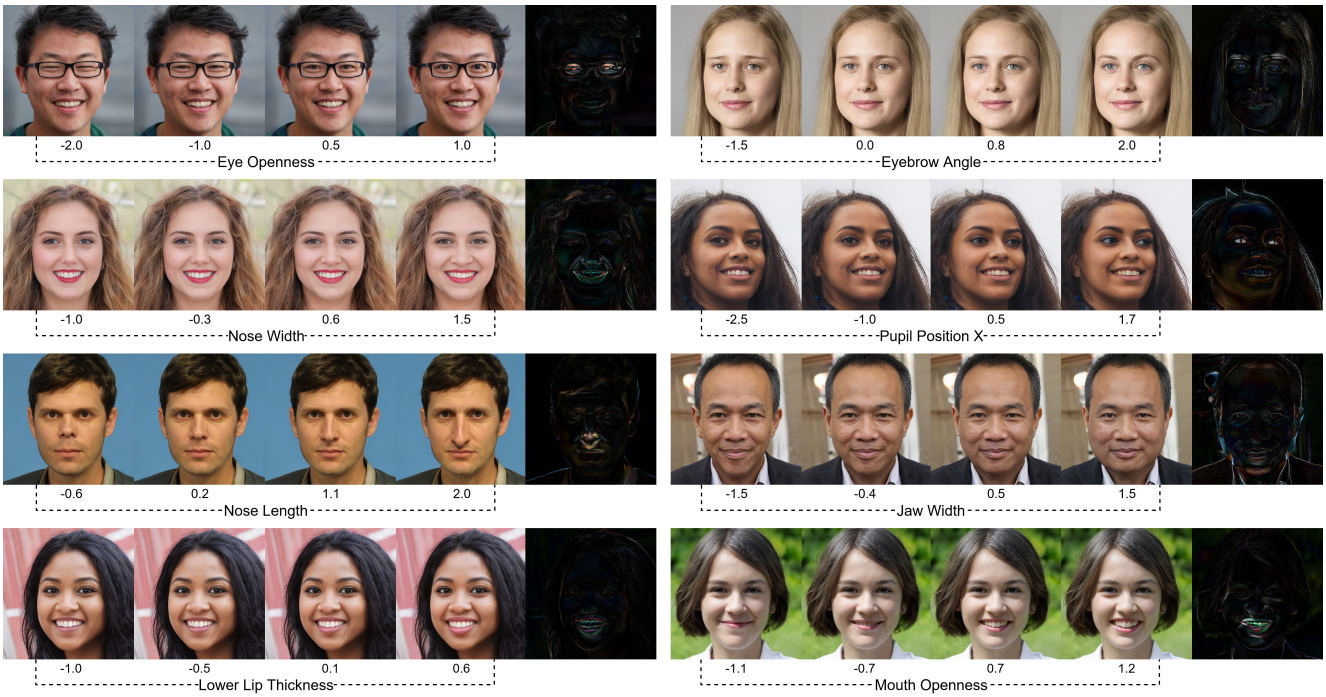

Figure 3. Additional progressive editing examples with UP-FacE. Each row shows a longer edit sequence, providing more examples than the corresponding figure in the main paper.

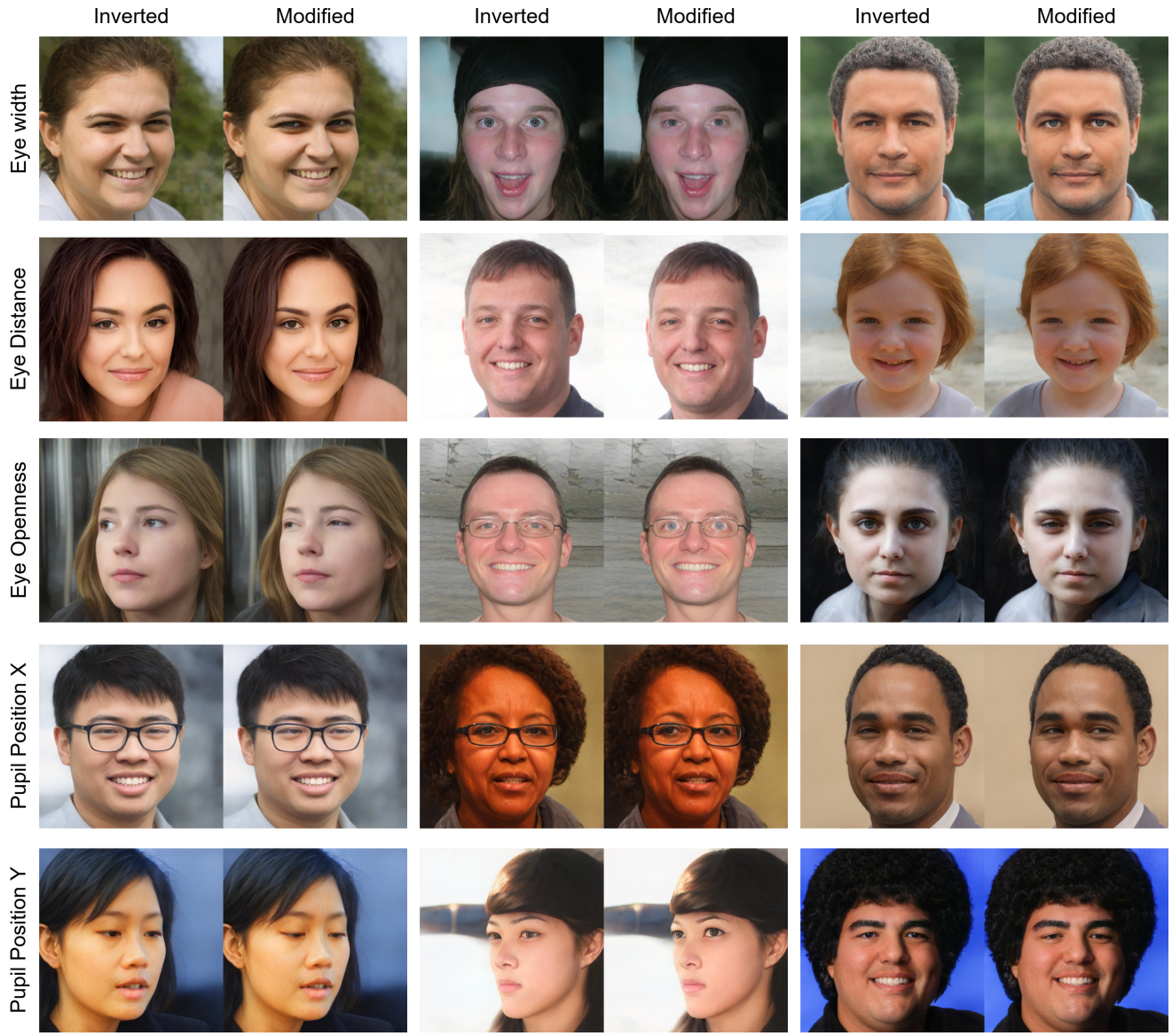

Figure 4. Additional real-image edits for our semantic features. Each pair shows an original inverted face (left) and the edited result (right).

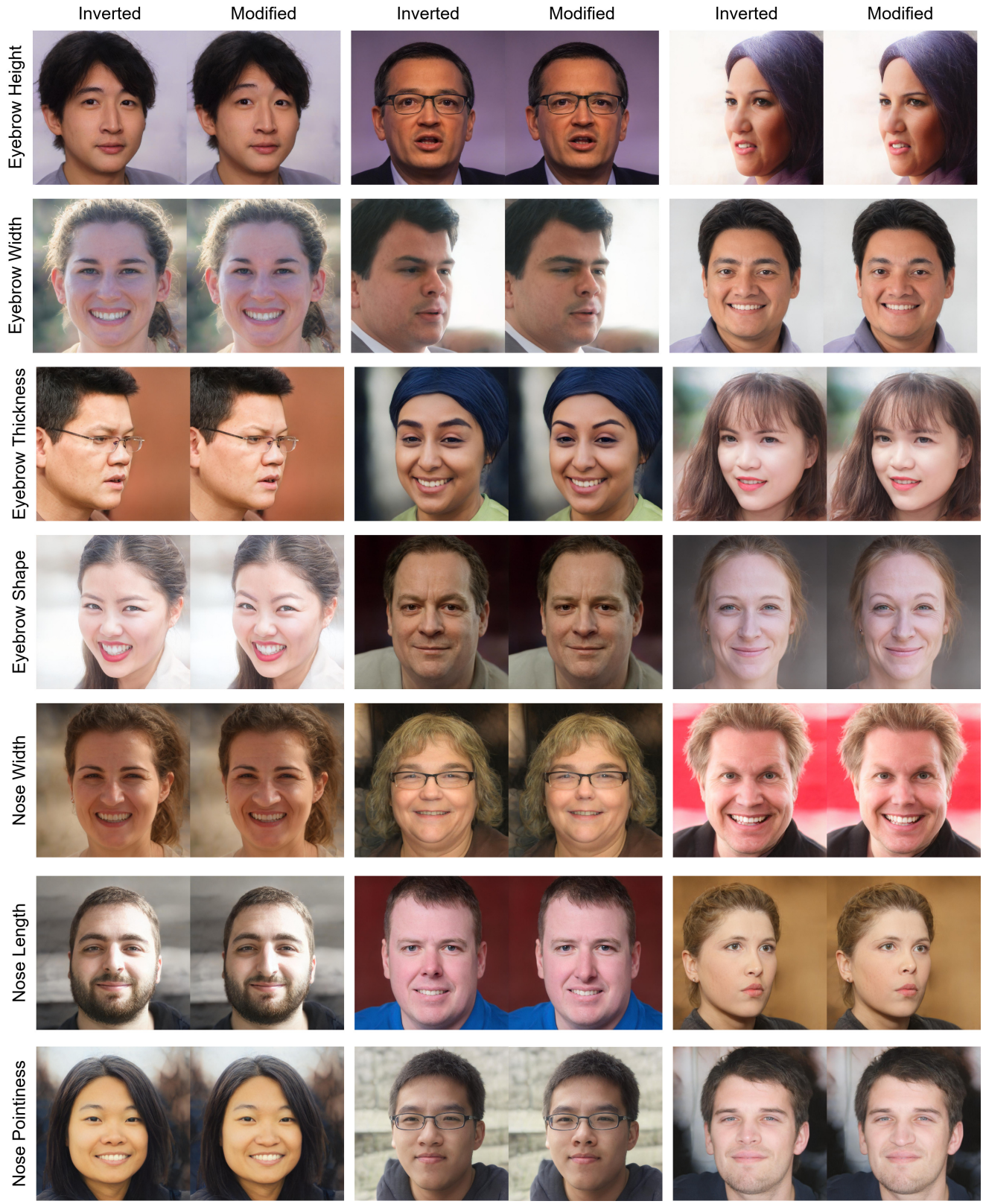

Figure 5. Additional real-image edits for our semantic features. Each pair shows an original inverted face (left) and the edited result (right).

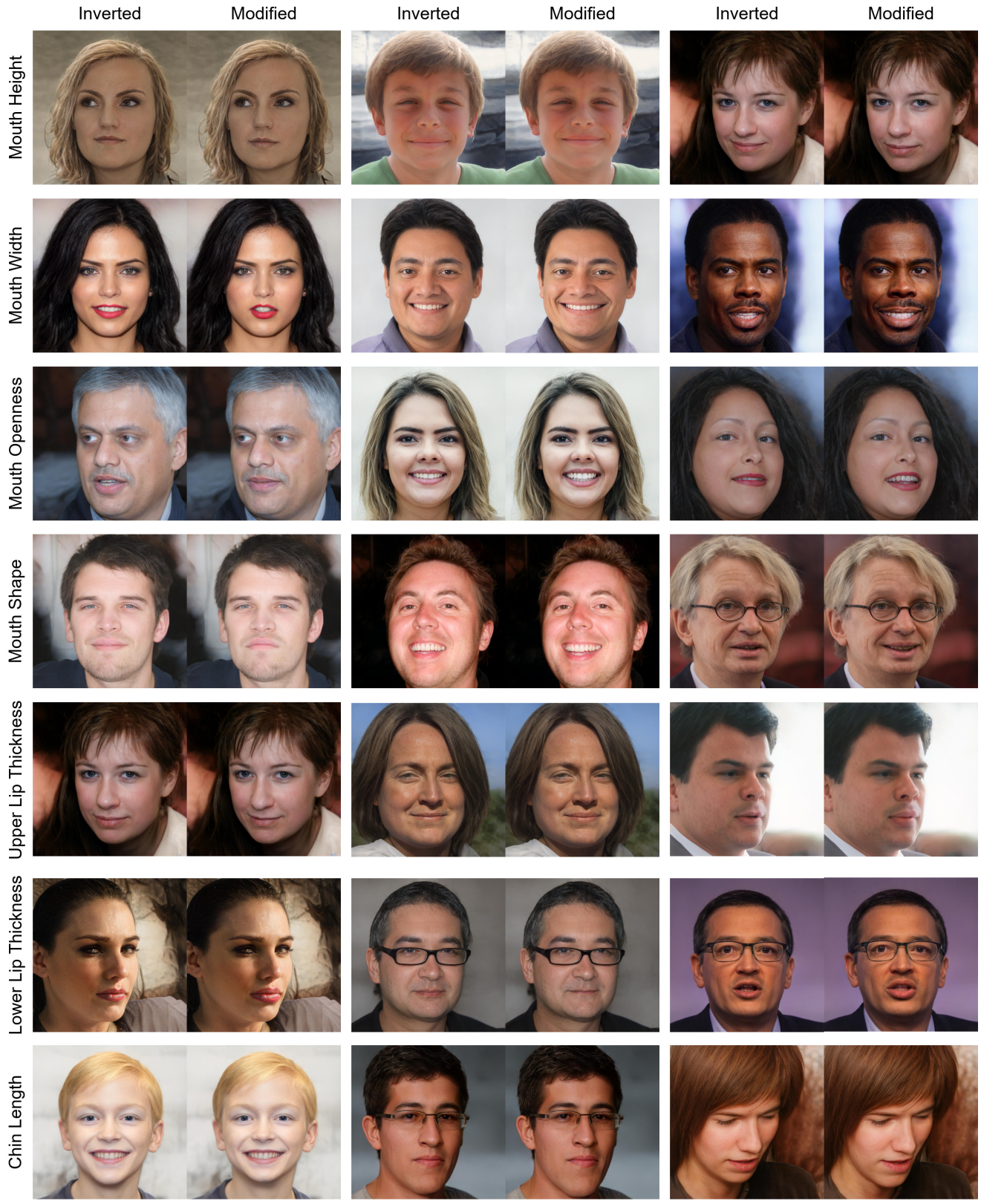

Figure 6. Additional real-image edits for our semantic features. Each pair shows an original inverted face (left) and the edited result (right).

Figure 7. Additional real-image edits for our semantic features. Each pair shows an original inverted face (left) and the edited result (right).