ChunkyGAN - Supplementary Material

Adéla Šubrtová*[1], David Futschik*[1], Jan Čech[1],
Michal Lukáč[2], Eli Shechtman[2], and Daniel Sýkora[1]

[1] Czech Technical University in Prague,
Faculty of Electrical Engineering, Czech Republic
{subrtade,futscdav,cechj,sykorad}@fel.cvut.cz
[2] Adobe Research, USA
{lukac,elishe}@adobe.com

In this document, we present additional experiments that provide further insight and evaluations. Namely, in Sections 1 and 2, we quantitatively compare with additional competing methods: three encoders pSp [7], e4e [9], ReStyle [2] and two optimization-based methods Pivotal Tuning [8] and HyperStyle [3]. We show more challenging examples and qualitative results of all tested methods. Section 3 shows the automatic segmentation and corresponding images for both cases of our method, with and without the regularization. Finally, in Section 4 we demonstrate additional examples of interactive image editing and application of our method to the image interpolation task.

1 Projection Fidelity

Projection	LPIPS	Identity	L_2
$\mathcal{W}$	0.4190 ± 0.0363	0.1745 ± 0.1328	0.0725 ± 0.0699
Ours in $\mathcal{W}$	0.3697 ± 0.0396	0.1384 ± 0.1117	0.0481 ± 0.0289
$\mathcal{W}^+$	0.3675 ± 0.0387	0.1195 ± 0.1047	0.0436 ± 0.0623
Ours in $\mathcal{W}^+$	**0.3194 ± 0.0365**	0.0937 ± 0.0855	0.0207 ± 0.0151
Ours in $\mathcal{W}^+$ reg.	0.3330 ± 0.0350	**0.0894 ± 0.074**	0.0217 ± 0.0130
$\mathcal{S}$	0.3577 ± 0.0397	0.1070 ± 0.0965	0.0328 ± 0.0188
Ours in $\mathcal{S}$	0.3572 ± 0.0401	0.1053 ± 0.0928	0.0319 ± 0.0187
e4e [9]	0.4444 ± 0.0418	0.1912 ± 0.1343	0.0468 ± 0.0165
pSp [7]	0.4433 ± 0.0418	0.1706 ± 0.1182	0.0351 ± 0.0135
ReStyle [2] - 5 iters	0.4444 ± 0.0430	0.1900 ± 0.1318	0.0433 ± 0.0162
Pivotal Tuning [8]	0.3332 ± 0.0353	0.0936 ± 0.0616	**0.0135 ± 0.0071**
HyperStyle [3] - 5 iters	0.4297 ± 0.0404	0.1420 ± 0.1003	0.0247 ± 0.0115

Table 1. Projection fidelity (extended)—losses were measured between the projected and the original image for each of the projection methods. Each cell reports the loss averaged over the CelebA subset along with the standard deviation.

The experiment measures the average LPIPS, Identity, and L_2 losses between the original and inverted images on CelebA subset of 100 images. Table 1

*joint first authors

is extended by five rows with other methods compared to Table 1 in the paper. Notably, all fast encoder-based approaches e4e [9], pSp [7], and ReStyle [2] produce lower fidelity images. In the case of Pivotal Tuning [8] we started refining the StyleGAN2 model from $\mathcal{W}^+$ codes as pivots. Our method performs better when measuring LPIPS and Identity. Pivotal Tuning is superior by a small margin only in the case of L_2 metric, which is known to be less correlated with the human perception. Moreover, the major drawback of Pivotal Tuning is that it requires to store the entire StyleGAN2 model for each image together with the corresponding latent code. HyperStyle [3] achieves better results than encoder methods but does not preserve the identity as faithfully as our method or Pivotal Tuning. See qualitative results in Fig. 1–Fig. 6 to compare differences among the methods visually.

2 Editability

	(a)				(b)			
	gender	smile	age	beard	gender	smile	age	beard
$\mathcal{W}$	0.169	0.022	0.07	0.279	0.249	0.18	0.191	0.328
$\mathcal{W}^+$	0.209	0.02	0.095	0.296	0.256	0.128	0.171	0.325
Ours in $\mathcal{W}^+$	0.298	0.049	0.151	0.312	0.325	0.125	0.203	0.333
Ours in $\mathcal{W}^+$ reg.	0.126	**0.018**	0.069	0.091	0.169	**0.099**	**0.129**	**0.144**
e4e [9]	**0.088**	0.024	**0.054**	0.239	0.26	0.242	0.245	0.351
pSp [7]	0.153	0.026	0.126	**0.074**	0.282	0.223	0.258	0.248
ReStyle [2]- 5 iters	0.097	0.030	0.081	0.213	0.417	0.409	0.399	0.453
Pivotal Tuning [8]	0.135	0.037	0.089	0.329	0.237	0.176	0.200	0.388
HyperStyle [3]- 5 iters	0.107	0.12	0.135	0.107	**0.15**	0.163	0.166	0.157

Table 2. Identity preservation during editing (extended)—identity loss was computed between the projected and the edited images (a), and between the original and the edited images (b).

We tested the five extra methods for identity preservation during editing. Table 2 extends the same table in the paper. The calibrated edits were made and the angular identity loss was measured. From Table 2 it is apparent that with regularization our method compares very favorably when examining the projected and edited images (a). Nevertheless, it outperforms by a large margin all competitors when comparing the *original* and edited images (b) for smile, age and beard edits. This is caused by the fact that the projection quality of previous approaches is not very faithful to the original image as can be seen quantitatively in Table 1 and qualitatively in Fig. 7. Pivotal Tuning gives a fair inversion quality, however, it can be seen that the facial masks are blurred and the subtle bindis were not reconstructed at all. The editing for Pivotal Tuning performs well but in the edited images, the identity is not well preserved.

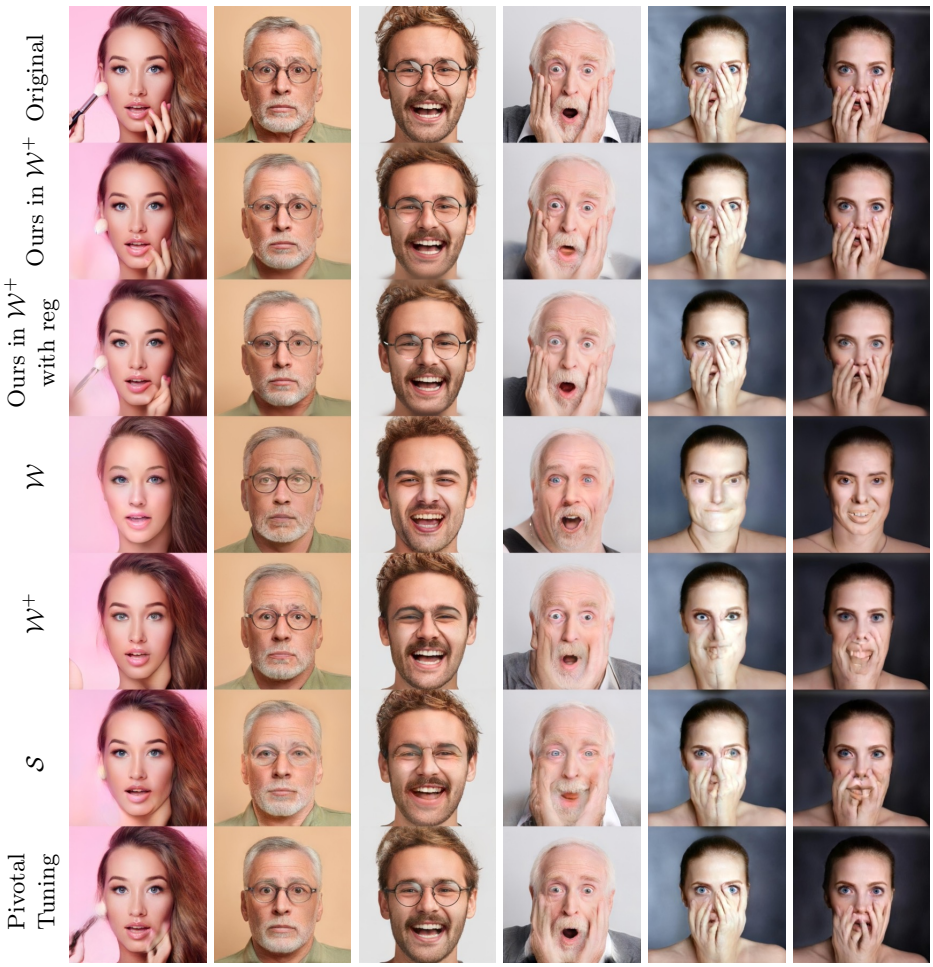

Fig. 1. Qualitative assessment of projection fidelity on challenging examples using **optimization-based methods** ($\mathcal{W}$ [6], $\mathcal{W}^+$ [1], $\mathcal{S}$-space [10], and Pivotal Tuning [8]). The images with the face occluded by hands are especially difficult to project for existing methods. The best results are produced by our approach. Pivotal Tuning faithfully projects the images with glasses (columns 2 and 3) but hands in columns 4 and 6 do not look very realistic. Source images: Adobe Stock.

Fig. 2. Qualitative assessment of projection fidelity on challenging examples using **encoder-based methods** (e4e [9], pSp [7], ReStyle [2], HyperStyle [3], and HyperInverter [5]). Our method consistently outperforms encoder-based techniques which struggle namely with the projection of the face occluded by hands in columns 4, 5, and 6. All approaches manage to reproduce glasses in columns 2 and 3 but only HyperStyle and our method correctly generate reflections in the column 2. Source images: Adobe Stock.

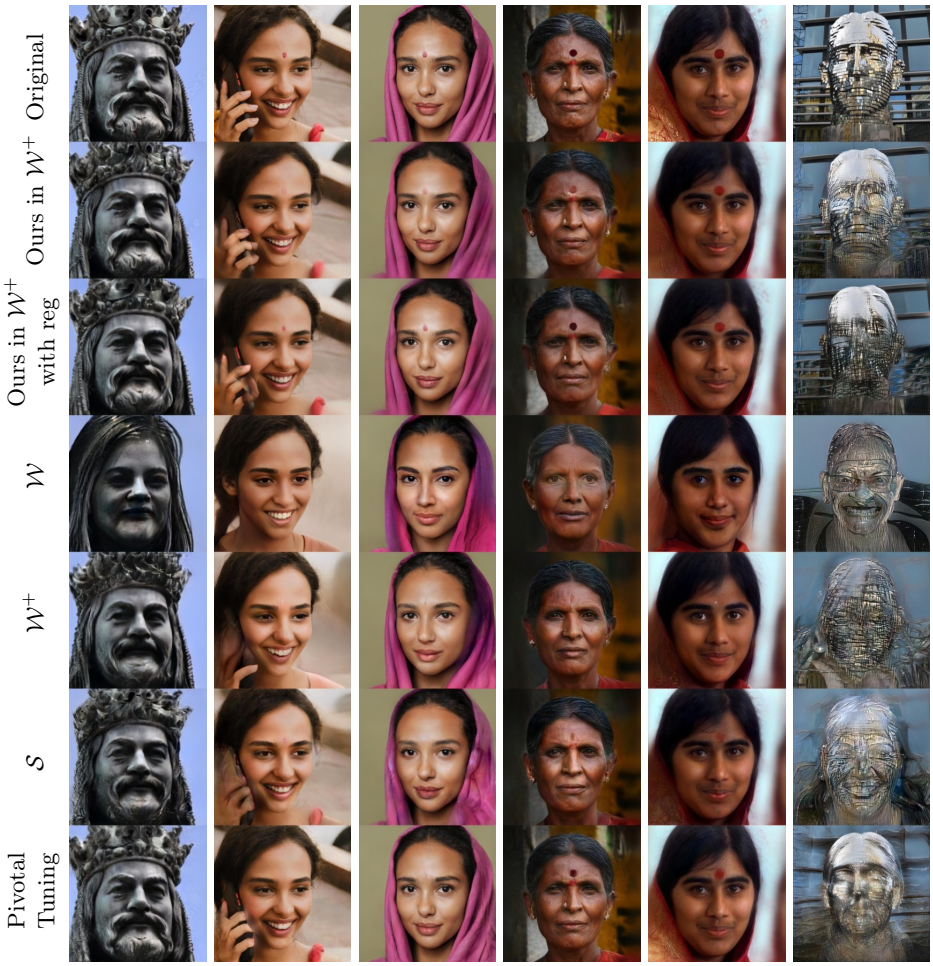

Fig. 3. Qualitative assessment of projection fidelity on challenging examples using **optimization-based methods** ($\mathcal{W}$ [6], $\mathcal{W}^+$ [1], $\mathcal{S}$-space [10], and Pivotal Tuning [8]). Our method enables to reproduce fine details such as bindi in columns 2, 3, 4, and 5. Pivotal Tuning is also able to faithfully synthesize bindi but only when there is a high contrast between the skin tone and the color of bindi. The statue in the column 6 is especially challenging to project. The best results are produced using Pivotal Tuning and our method without regularization. Source images: columns 1–5: Adobe Stock, column 6: Jindich Nosek (NoJin).

Fig. 4. Qualitative assessment of projection fidelity on challenging examples using **encoder-based methods** (e4e [9], pSp [7], ReStyle [2], HyperStyle [3], and HyperInverter [5]). In contrast to our approach encoder-based methods fail to reproduce realistic bindi (columns 2, 3, 4, and 5). HyperStyle and HyperInverter manage to preserve the identity well for in-domain images. Source images: columns 1–5: Adobe Stock, column 6: Jindich Nosek (NoJin).

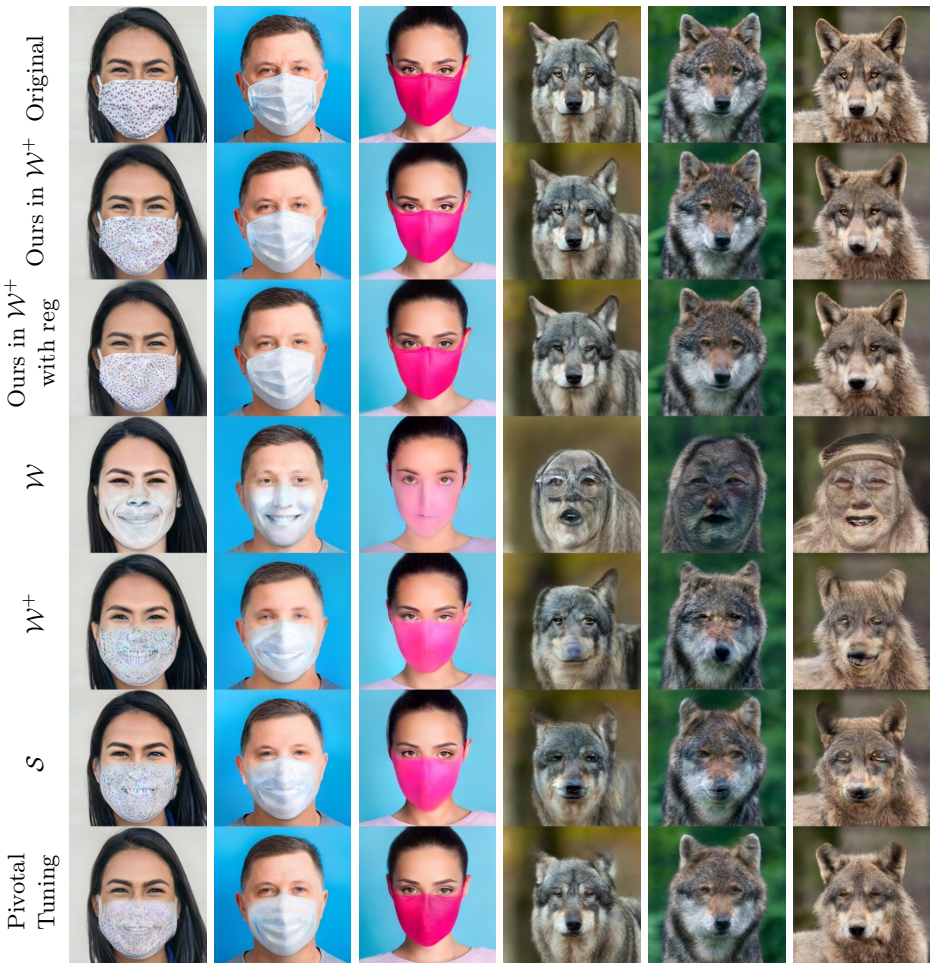

Fig. 5. Further comparison on images which are challenging to invert accurately using existing **optimization-based approaches** ($\mathcal{W}$ [6], $\mathcal{W}^+$ [1], $\mathcal{S}$-space [10], and Pivotal Tuning [8]). Although Pivotal Tuning manage to reconstruct face masks relatively well their boundaries are blurry. $\mathcal{S}$-space method can generate wolves (columns 5 and 6) to some extent but they are still far from the original images. Source images: Adobe Stock.

Fig. 6. Further comparison on images which are challenging to invert accurately using **encoder-based methods** (e4e [9], pSp [7], ReStyle [2], HyperStyle [3], and HyperInverter [5]). The face masks and the out-of-domain images are especially challenging for the encoder-based approaches. HyperStyle and HyperInverter produce correct shape for the wolfs in columns (4, 5, and 6) but the details are not realistic. Source images: Adobe Stock.

Concerning limitations of our method, they occur in case of extreme edits that significantly change the geometry of the image, such as yaw. In that case, there is a disparity between segments and the projections, thus visible seam artifacts appear, see Fig. 8. We shortly discussed in the main paper in Section 6 possible future options to resolve the issue.

3 Regularization

In Table 2, it is seen that the regularization has a positive impact on the identity preservation during editing. We believe the reason is that the non-regularized projection may generate unrealistic images with codes far from learned manifold of the latent space. The regularization encourages the codes to be closer to the mean, producing in-domain images for which the editing by latent code manipulation along pre-trained semantic directions works better. The effect of the regularization is demonstrated in Fig. 9. In both cases, the composed image is very faithful to the original, and the composed images are hard to distinguish. However, the component images are notably more realistic when the regularization is on. Out-of-domain example is shown in Fig. 10.

4 Additional Applications

Our approach can be used for image interpolation task (see Fig. 11). In this application two estimated latent codes in each segment are linearly interpolated to produce partial inbetween image which is then composed with other inbetween images using Laplacian pyramid [4]. The latent codes of the segments were interpolated as well. Since our method preserves identity better and the corresponding facial features are naturally mapped on each other due to the interpolation in the latent space the resulting transition looks convincing.

In Fig. 12 we present an editing example produced using our method where a StyleGAN2 model trained on cars was used as a backbone. This example demonstrates that our approach is agnostic to the domain on which the model was trained. The only requirement for the used model is that it has a sufficient number of pre-trained directions for latent code manipulation.

As demonstrated in the main paper our method is flexible enough to make a projection of out-of-domain images, i.e., images that were not considered during the training of StyleGAN2 model. In Fig. 13 we edit various famous paintings using our method with StyleGAN2 model trained exclusively on photographs of real faces. Thanks to the accurate projection the subsequent edits look like if they were produced by a StyleGAN2 model trained on paintings.

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2StyleGAN: How to embed images into the StyleGAN latent space? In: Proceedings of IEEE International Conference on Computer Vision (2019)

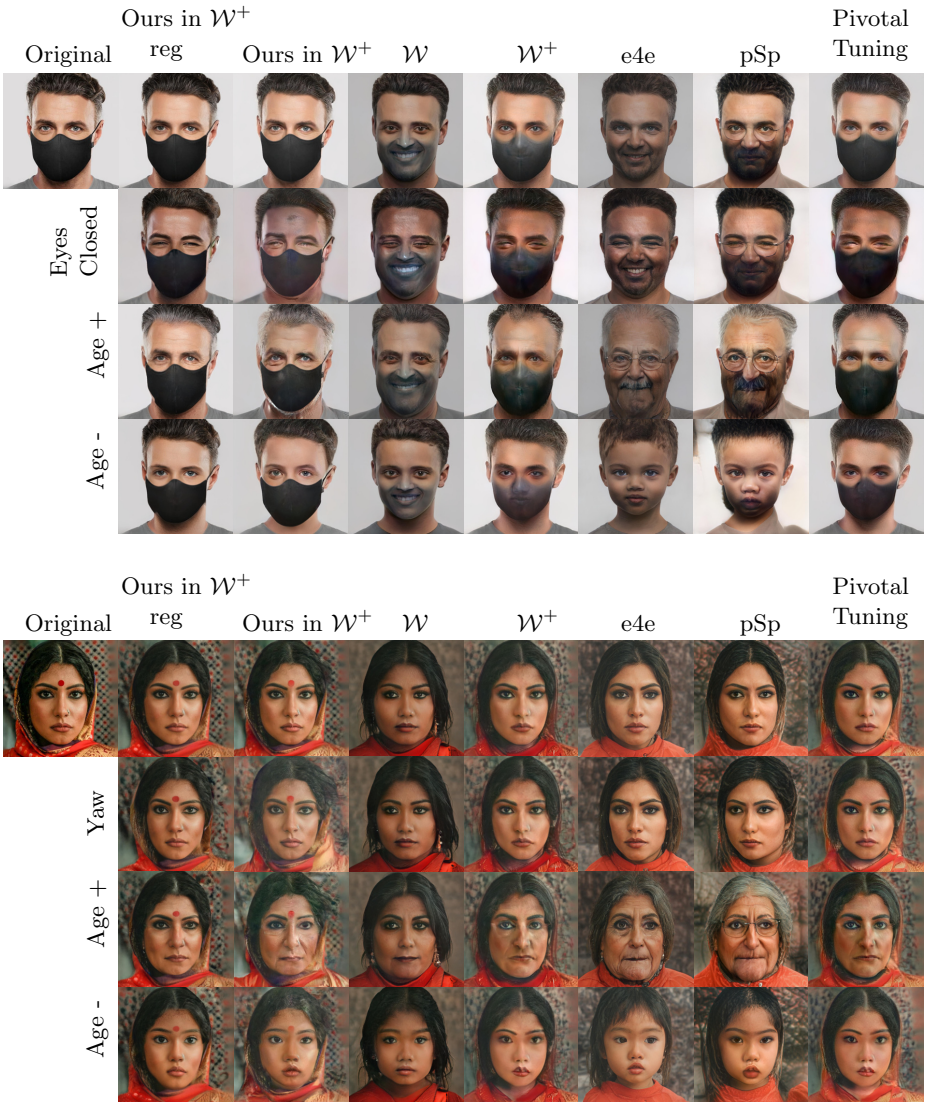

Fig. 7. Challenging global edits. The row besides the original image shows inverted images by all tested methods ($\mathcal{W}$ [6], $\mathcal{W}^+$ [1], e4e [9], pSp [7], and Pivotal Tuning [8]). Other rows display corresponding edited results along given semantic directions. For our methods, the editing was done simply by manipulating the latent codes the same for all the segments. The results of the inversion and editing are fully automatic. No manual adjustments or post-processing were performed. Source images: Ayush Kejriwal (bindi), BlochWorld (face mask).

Fig. 8. Limitations of using our method to perform edits which change the geometry to a significant extent. In these examples, segment seams become visible for larger yaw changes without any explicit treatment of the segments. Source images: Adobe Stock (brush), Mingle Media TV (Kate Winslet), Neil Grabowsky / Montclair Film (Ethan Hawke), NASA/Aubrey Gemignani (Ryan Gosling).

(a) Regularization on ($\lambda_{reg} = 1$)

(b) Regularization off ($\lambda_{reg} = 0$)

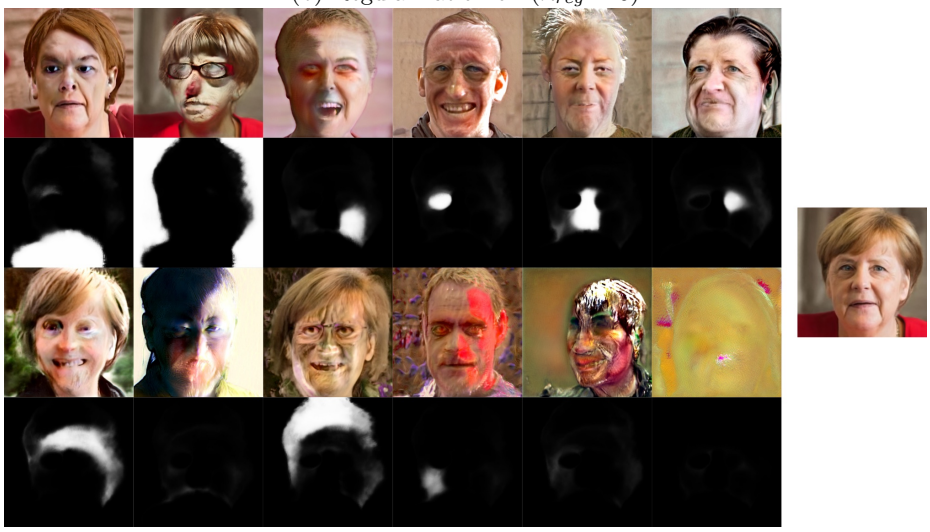

Fig. 9. Effect of regularization. The projected (composed) images are on the right. The left side depicts individual projections with the corresponding segmentation masks underneath. Source image: Raimond Spekking / CC BY-SA 4.0 (via Wikimedia Commons).

(a) Regularization on ($\lambda_{reg} = 1$)

(b) Regularization off ($\lambda_{reg} = 0$)

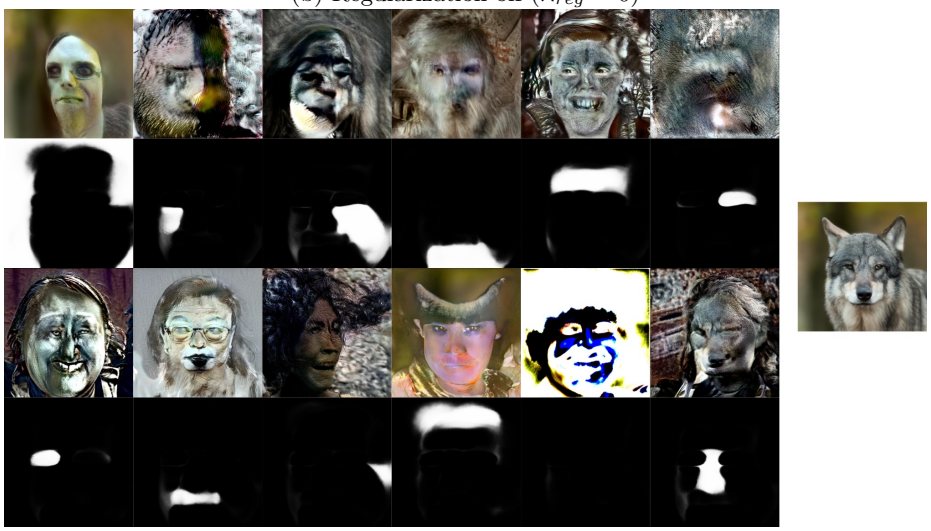

Fig. 10. Effect of regularization - out-of-domain example. The projected (composed) images are on the right. The left side depicts individual projections with the corresponding segmentation masks underneath. Source image: Adobe Stock.

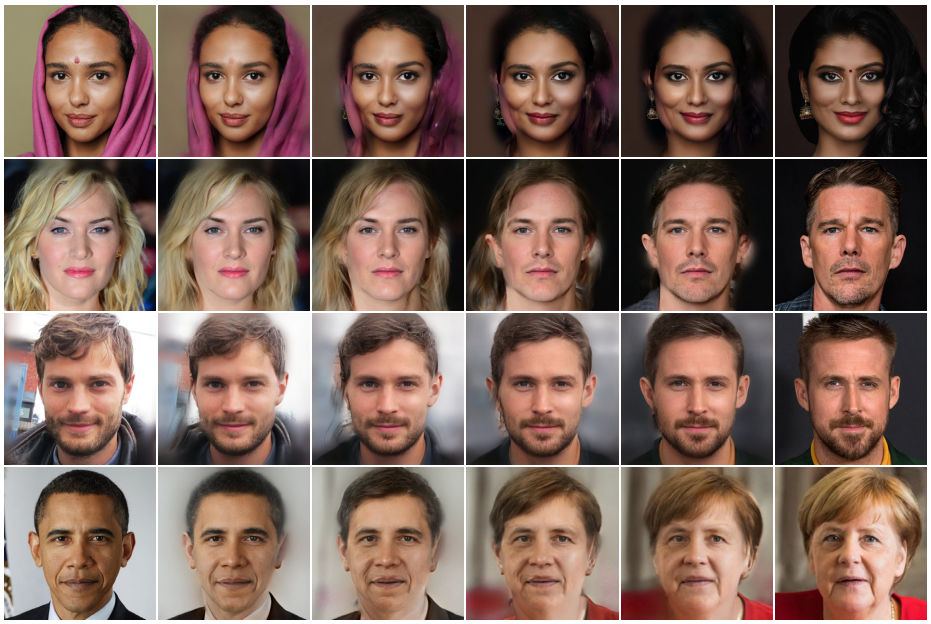

Fig. 11. Interpolation examples—our approach can be used to perform interpolation between two different identifies. The estimated latent code in each segment is linearly interpolated and the final image is then composed using Laplacian pyramid [4]. A key advantage here is that in our method identity is preserved better and thus the transition looks more believable. Source images: Adobe Stock (bindis), Mingle Media TV (Kate Winslet), Neil Grabowsky / Montclair Film (Ethan Hawke), katmtan (Jamie Dornan), NASA/Aubrey Gemignani (Ryan Gosling), Pete Souza (Barack Obama), Raimond Spekking / CC BY-SA 4.0 (via Wikimedia Commons) (Angela Merkel).

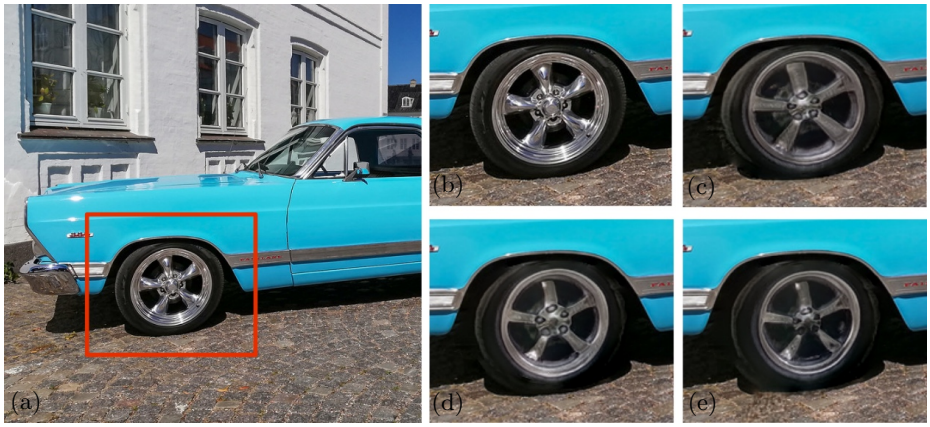

Fig. 12. Editing using our method based on StyleGAN2 model trained on photos with cars—original image (a), detail of the original image (b), local edits of wheel disc design (c–e). "The Blue Car NO Not my car" by munchflemming is licensed under CC BY 2.0.

2. Alaluf, Y., Patashnik, O., Cohen-Or, D.: ReStyle: A residual-based StyleGAN encoder via iterative refinement. In: Proceedings of IEEE International Conference on Computer Vision. pp. 6711–6720 (2021)
3. Alaluf, Y., Tov, O., Mokady, R., Gal, R., Bermano, A.H.: HyperStyle: StyleGAN inversion with hypernetworks for real image editing. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 18511–18521 (2022)
4. Burt, J.R., Adelson, E.H.: A multiresolution spline with application to image mosaics. ACM Transactions on Graphics **2**(4), 217–236 (1983)
5. Dinh, T.M., Tran, A.T., Nguyen, R., Hua, B.S.: HyperInverter: Improving stylegan inversion via hypernetwork. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 11389–11398 (2022)
6. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 8107–8116 (2020)
7. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: A stylegan encoder for image-to-image translation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 2288–2296 (2021)
8. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. In: arXiv. No. 2106.05744 (2021)
9. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for StyleGAN image manipulation. ACM Transactions on Graphics **40**(4), 133 (2021)
10. Wu, Z., Lischinski, D., Shechtman, E.: Stylespace analysis: Disentangled controls for StyleGAN image generation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 12863–12872 (2021)

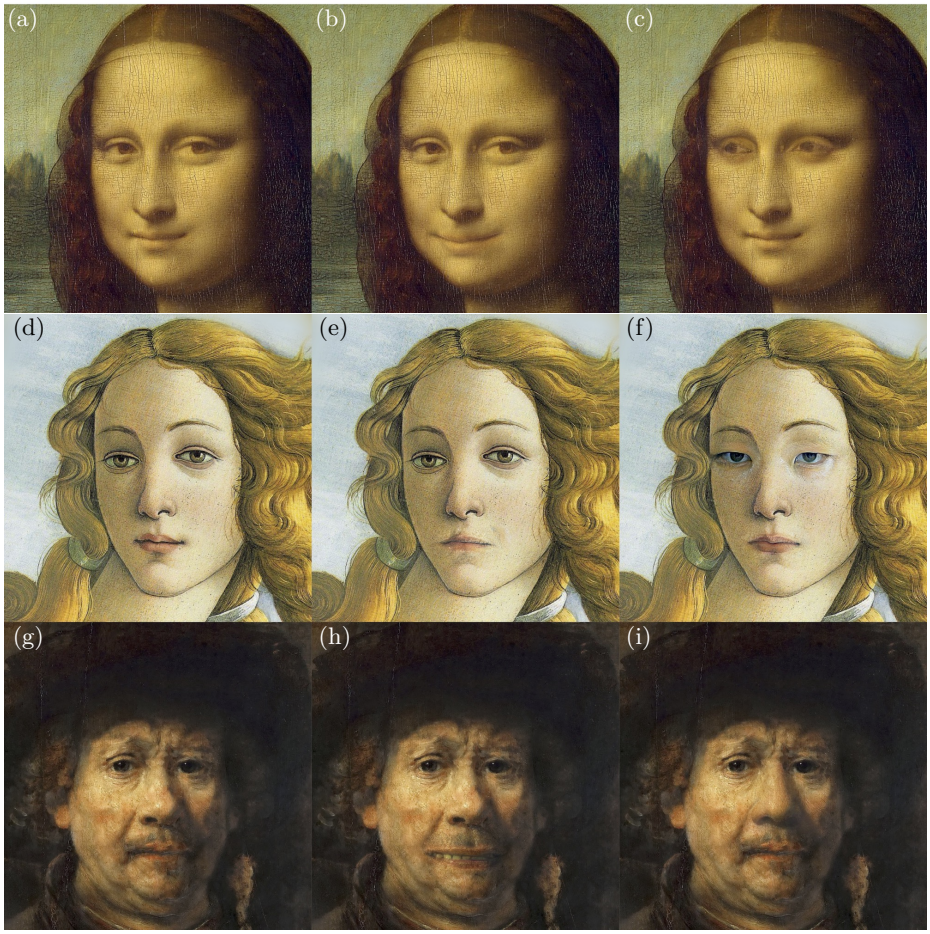

Fig. 13. Edits performed on famous painting using our approach with StyleGAN2 model trained on real faces—original Da Vinci's Mona Lisa (a), more pronounced smile (b), change in the gaze direction (c), original Botticelli's The Birth of Venus (d), change in the mouth expression (e), different shape of eyes (f), original Rembrandt's Little Self-portrait (g), changing mouth expression (h), different shape of the nose (i).