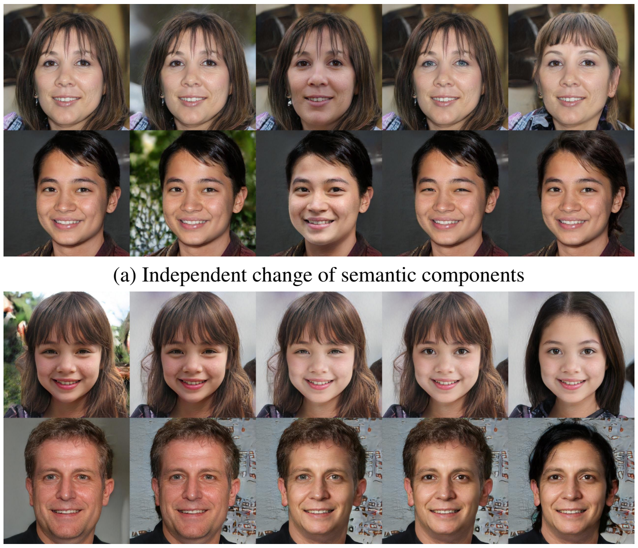
VerA: Versatile Anonymization Applicable to Clinical Facial Photographs

Majed El Helou*[1], Doruk Cetin*[2], Petar Stamenkovic[1], Niko Benjamin Huber[2], Fabio Zünd[1]
[1]ETH Zürich, Switzerland, [2]Align Technology, Switzerland
{majed.elhelou,petars,fabio.zund}@ethz.ch, {dcetin,nhuber}@aligntech.com

Abstract

We present extended results illustrating the control of our image generator, both in terms of semantic and high-level generative control. We additionally propose extended anonymization evaluation for the different problem settings. Namely, further results on standard single-image anonymization, clinical single-image anonymization, as well as the paired counterparts where two images of the same person need to be anonymized consistently. We also present illustrative examples of full-image anonymization and more evaluation on downstream utility. Lastly, we provide an ablation study of our proposed mirroring contrastive learning and the projection heads we learn on top of the pretrained high-level encoders. We add the table of contents below for more convenience in navigating between the sections. All the supplementary results support what we present in our main manuscript.

Table of Contents



(a) Independent change of semantic components

(b) Same as (a), with accumulation of changes from left to right

Figure 1. Semantic control with changes applied (a) independently, and (b) cumulatively from left to right. The leftmost image is the original, and from left to right we change both the structure and texture of: *background, face, eyes* and *hair*.

1. Extended results of controllable synthesis

1.1. Semantic generative image control

We illustrate the disentangled semantic control capabilities of our generator in Fig. 1. We show two examples in Fig. 1a where each column has a different semantic change relative to the leftmost column. In Fig. 1b, the changes are accumulated from left to right, modifying in order the background, face, eyes and hair. This semantic control is due to the architecture components from SemanticStyleGAN [12] that extends on StyleGAN2 [7], and we show here that the high-level control that we achieved with our training does not block the semantic control.

1.2. High-level generative image control

We show the high-level attribute control that our model achieves in Fig. 2 flexibly on pose and age. We also

[*] These authors contributed equally to this work.

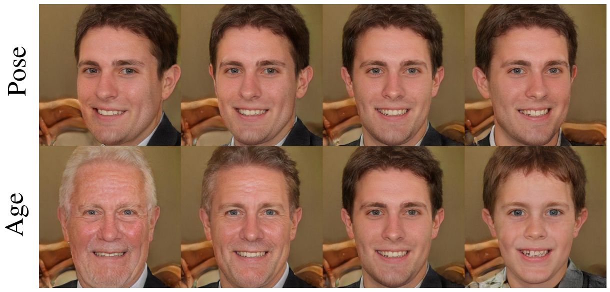
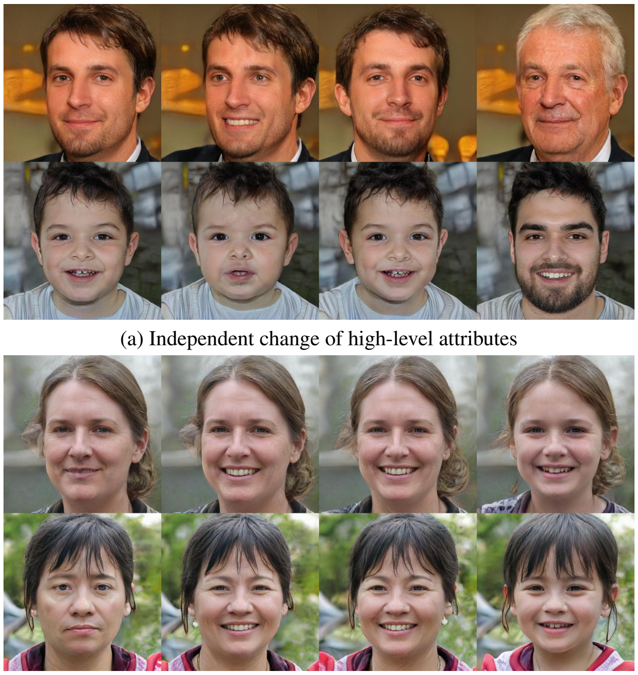


Figure 2. Illustration of high-level changes, on the same person. The top row modifies the *pose* gradually, and the bottom row modifies *age* gradually. All other attributes remain unchanged.



(a) Independent change of high-level attributes

(b) Same as (a), with accumulation of changes from left to right

Figure 3. High-level attribute control, with changes applied (a) independently, and (b) cumulatively from left to right. The leftmost image is the original, and from left to right we change: *expression, orientation*, and *age*.

show independent modifications that we apply in expression, pose, and age in Fig. 3a. Fig. 3b shows similar results but with the accumulation of high-level changes from left to right, sequentially altering the person's expression, orientation, and age. Lastly, we show in Fig. 4 that even within pose we can disentangle yaw and pitch just with PCA. By performing a PCA decomposition over the pose latent, we can move in the direction of one component to modify yaw, and the other component for controlling pitch. As shown in the top part, we can linearly control the yaw (top left linear curves) without affecting the pitch (top right flat curves).

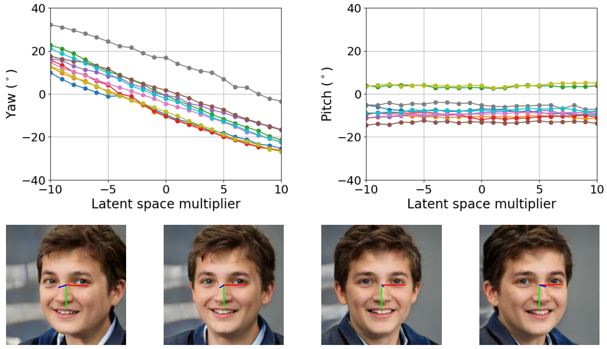


Figure 4. We linearly vary the multiplier (x-axis) that shifts the latent vector in the pose latent space along the yaw direction obtained from PCA and observe with a 6DRepNet [3] the resulting yaw (in the top left plot) and pitch (in the top right plot). We observe that the yaw linearly varies with our linear shift consistently across all images, while the pitch remains stable, proving the high-quality of our high-level attribute disentanglement. The bottom part shows a corresponding sample image with varying yaw.

2. Extended anonymization evaluation

2.1. Standard single-image anonymization

We provide extended benchmarking results of standard single-image anonymization in Fig. 5 on the CelebA-MaskHQ [9] test set and in Fig. 6 on the FFHQ [6] test set. We compare against the two most commonly referenced baselines CIAGAN [11] and FIT [2], and the three most recent state-of-the-art methods that have public code available; DP2 [5], RiDDLE [10] and FALCO [1]. Our VerA results are the most photorealistic, consistently de-identifying the person, even on this setting of standard single-image anonymization.

We make a note regarding the results of FALCO [1]. As mentioned in our main text, FALCO performs an adaptive normalization that can lead to washed out images, or to odd color artifacts if toggled off. We follow the authors' default setting and leave it activated in all our experiments. We illustrate this normalization's effect in Fig. 7.

2.2. Clinical single-image anonymization

We provide further benchmarking results on clinical single-image anonymization in Fig. 9, using images from our test set in FFHQ [6]. We compare against the same set of methods, and include our clinical anonymization results that preserve the mouth, eyes, and nose, respectively. We also provide extended quantitative evaluation on semantic preservation, conducted on FFHQ [6], in Table 1. All results support the same claims we make in our main manuscript. We further provide example results of competing methods, to which we add our own blending procedure in Fig. 8. Other methods do not directly tackle clinical anonymization

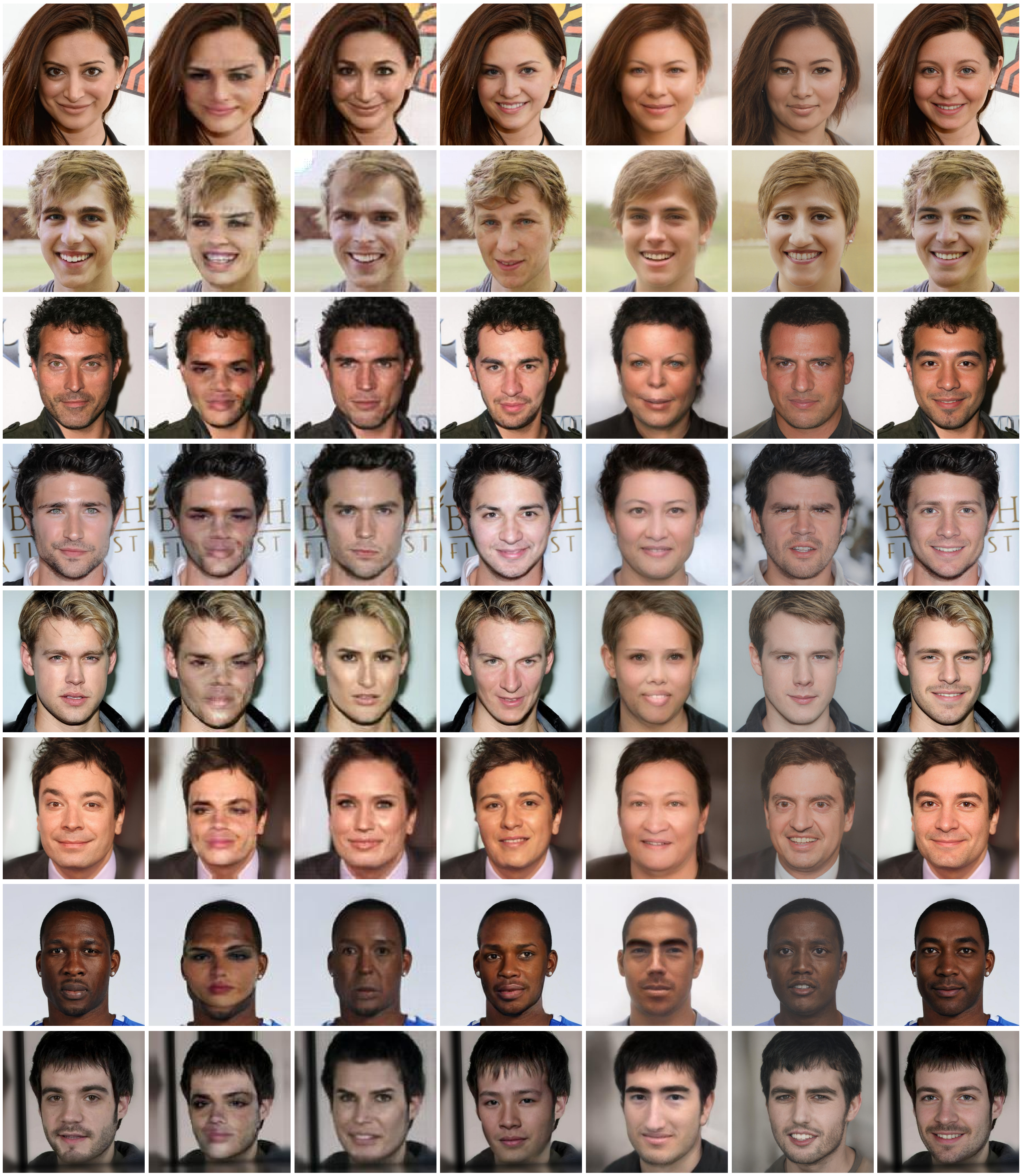


Figure 5. Extensive qualitative evaluation results on the *standard single-image* anonymization, benchmarking against the two most commonly referenced anonymization methods and the four most recent state-of-the-art anonymization approaches, on CelebAMaskHQ [9] test samples. Our proposed VerA achieves good photorealistic results consistently while de-identifying the input image, outperforming the best baselines even on this standard (non clinical) single-image anonymization task.

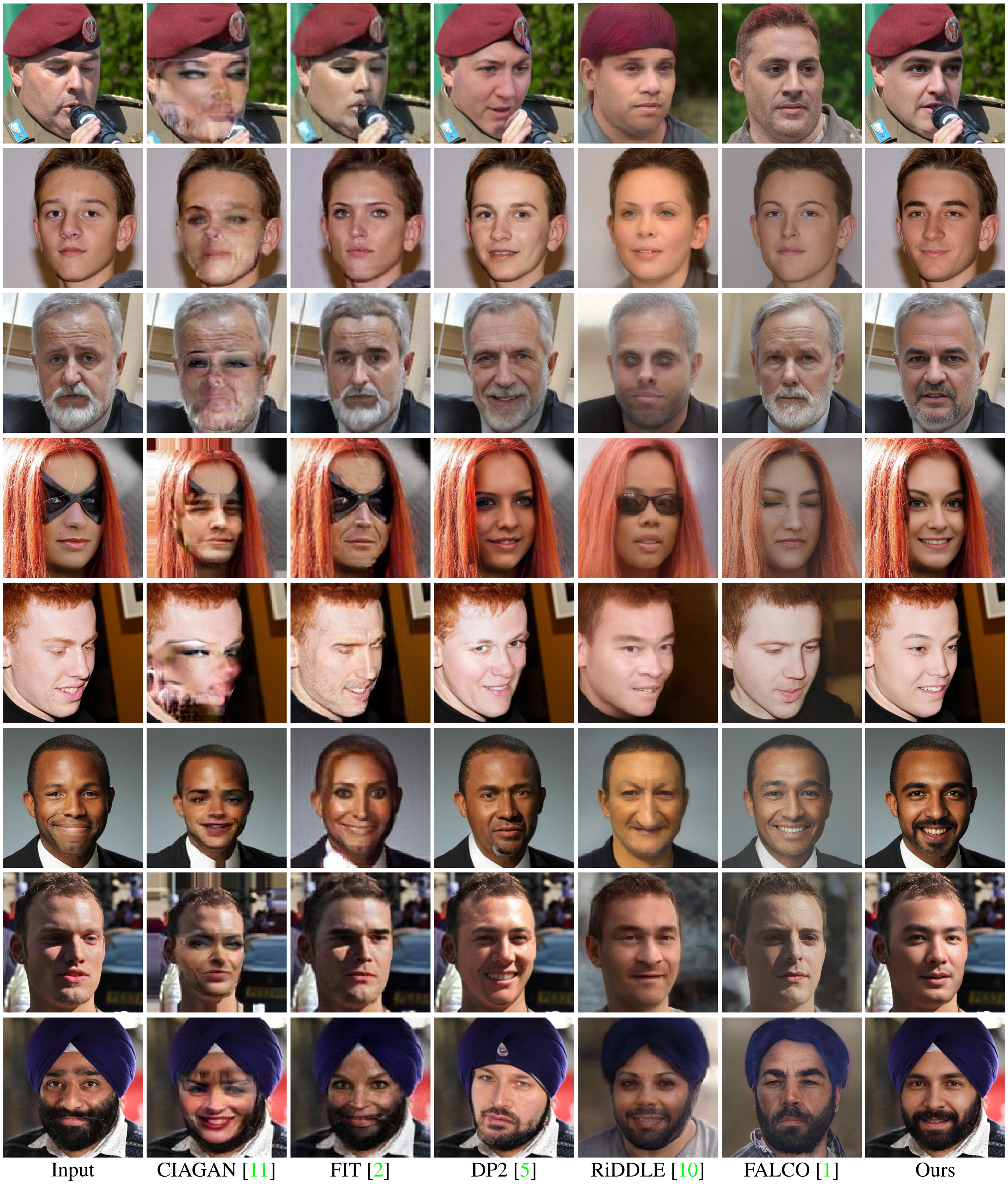


Figure 6. Extensive qualitative evaluation results on the *standard* single-image anonymization, similar to Fig. 5, but performed here on FFHQ [6] test samples. Thanks to semantic-aware inversion, VerA can robustly anonymize images with occluding objects and various accessories such as hats (first and last rows).

		ℓ_1 distance ↓			PSNR ↑			Semantic IoU ↑			Mean landmark offset ↓		
	Method	Mouth	Nose	Eyes	Mouth	Nose	Eyes	Mouth	Nose	Eyes	Mouth	Nose	Eyes
Standard	CIAGAN [11]	38.53	31.73	54.00	14.20	15.76	11.60	0.53	0.53	0.01	17.50	21.26	43.88
	FIT [2]	21.19	17.04	24.70	19.12	20.85	17.53	0.75	0.81	0.57	9.75	9.38	8.82
	DP2 [5]	40.72	32.95	49.16	13.58	15.05	12.05	0.52	0.62	0.23	29.82	32.45	28.13
	RiDDLE [10]	35.97	30.48	36.05	14.71	15.98	14.48	0.69	0.77	0.59	14.30	18.98	8.58
	FALCO [1]	33.63	27.01	34.93	15.25	17.07	14.70	0.63	0.76	0.54	18.22	17.40	9.08
	Ours	34.59	21.89	34.82	14.69	18.09	14.38	0.64	0.78	0.61	17.39	16.04	7.62
Clinical	Ours (mouth)	**0.22**	22.32	36.20	**54.42**	18.01	14.15	**0.90**	0.79	0.60	**8.13**	15.56	7.85
	Ours (nose)	34.63	**0.21**	36.05	14.67	**54.59**	14.20	0.65	**0.93**	0.60	16.29	**5.68**	7.68
	Ours (eyes)	34.91	22.17	**0.35**	14.61	18.03	**50.72**	0.64	0.78	**0.76**	17.84	16.58	**6.25**

Table 1. Semantic preservation results, in terms of content (ℓ_1, PSNR) and area (IoU, landmarks), evaluated on FFHQ [6]. We note two main observations: standard anonymization approaches destroy all semantic components that may need to be preserved in clinical images, and our clinical anonymization successfully preserves the desired component while also flexibly modifying non-blocked components as much as the baselines. Note: eye landmarks are key components in the alignment algorithm of FFHQ [6], which results in similar eye landmarks across images, thus the generally lower average landmark offset.

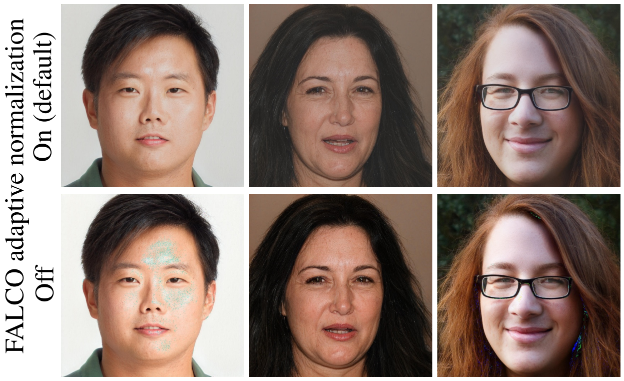


Figure 7. Illustration of the two settings of adaptive normalization in FALCO [1]. The authors' default corresponds to the top row, and can lead to washed out final images. If toggled off, this setting can lead to odd color artifacts like the blue in the first and last column (bottom right corner between the face and the hair).

Method	FID ↓		Bounding box ↑		Face detection ↑	
	FFHQ	CelebAHQ	MTCNN	Dlib	MTCNN	Dlib
CIAGAN [11]	109.92	93.46	0.80	0.88	0.90	0.90
FIT [2]	89.47	95.98	**0.91**	**0.94**	0.98	0.99
DP2 [5]	23.41	51.89	0.87	0.88	0.96	0.97
RiDDLE [10]	69.93	66.95	0.90	0.91	**1.00**	**1.00**
FALCO [1]	48.03	53.35	0.89	0.91	0.99	**1.00**
Ours	**13.79**	**51.60**	**0.91**	0.92	0.97	**1.00**

Table 2. Downstream utility evaluation for photorealism/diversity (FID [4]), bounding box IoU, and face detection rates (MTCNN [14], Dlib [8]), which we compute over FFHQ [6] test data. We achieve the best FID, and are on par with the best bounding box IoU and the best detection scores.

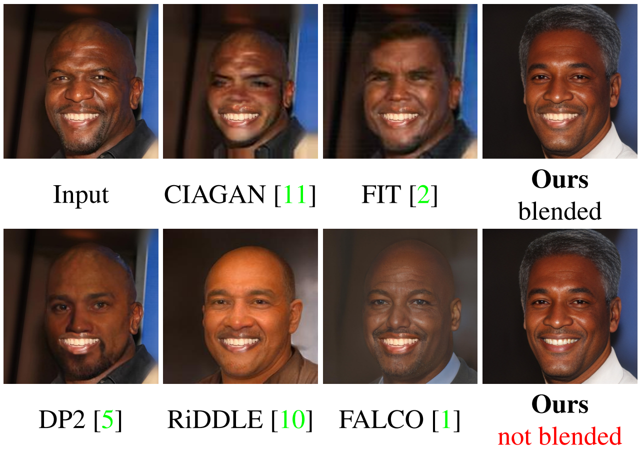


Figure 8. Best viewed **zoomed in**. We blend the *mouth* (of the same sample shown in Fig. 5 of the main manuscript) for competing methods, which yields less photorealistic results as competitors do not properly address semantic anonymization. We additionally show for illustration our own output without our region-of-interest blending.

with a semantic region-of-interest, and their results with our added blending are less photorealistic. We also show our own output when we do not perform our blending for illustration purposes.

2.3. Paired standard and clinical anonymization

We provide further examples of paired anonymization, in both the standard and clinical setting, and comparing to all benchmarks in Fig. 10 on a pair from the SiblingsDB dataset [13]. We additionally provide images at higher resolution for illustrating a paired clinical anonymization example in Fig. 12.

2.4. Full-image in-place anonymization

Fig. 14 shows numerous examples of full images that we anonymize using VerA. These examples serve as an illustra-

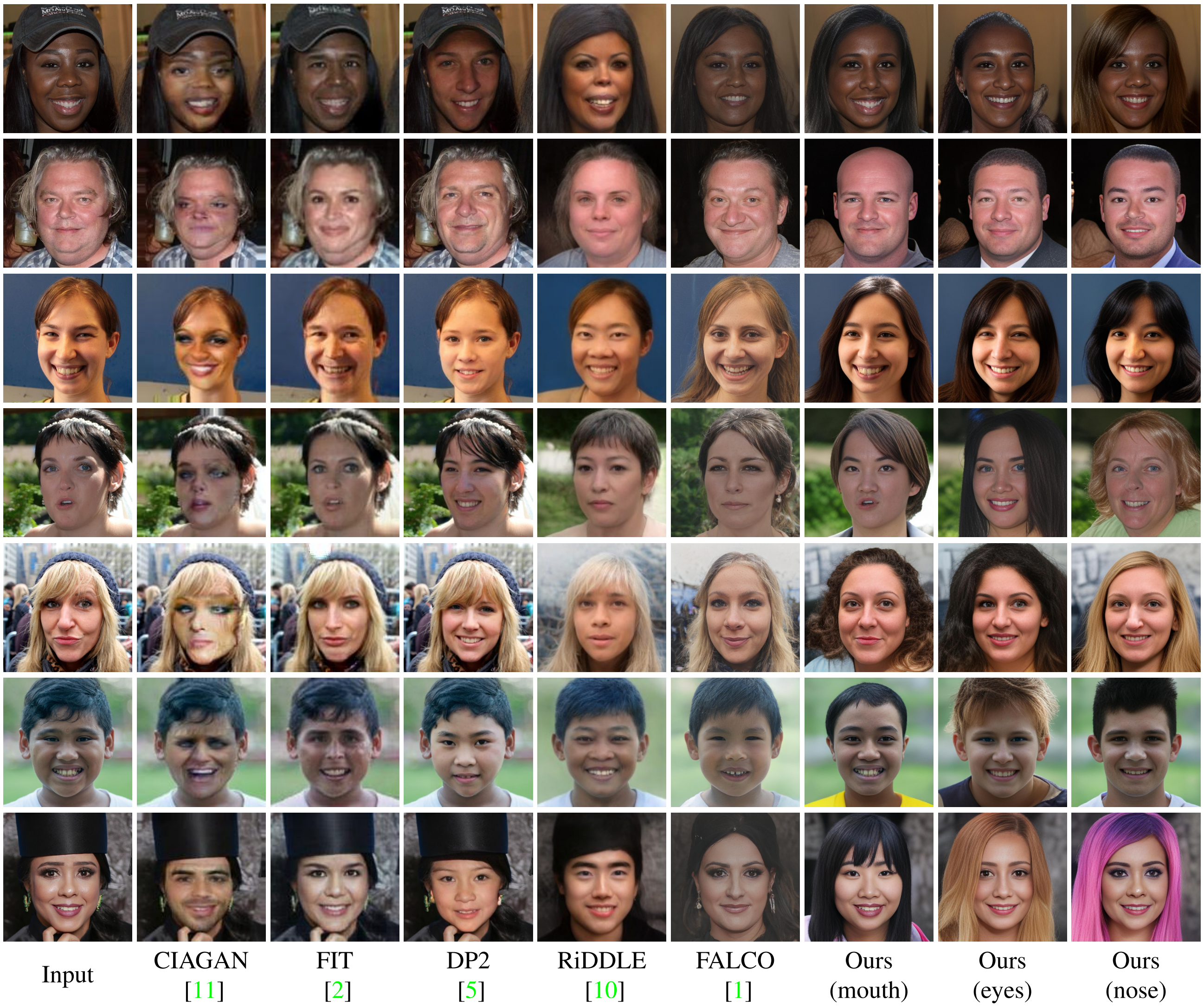


Figure 9. Extensive qualitative evaluation results on the *clinical* single-image anonymization, benchmarking against the two most commonly referenced anonymization methods and the four most recent state-of-the-art anonymization approaches, on FFHQ [6] test samples.

tion of the application to full-scene images, aside from the clinical use cases that our main manuscript focuses on.

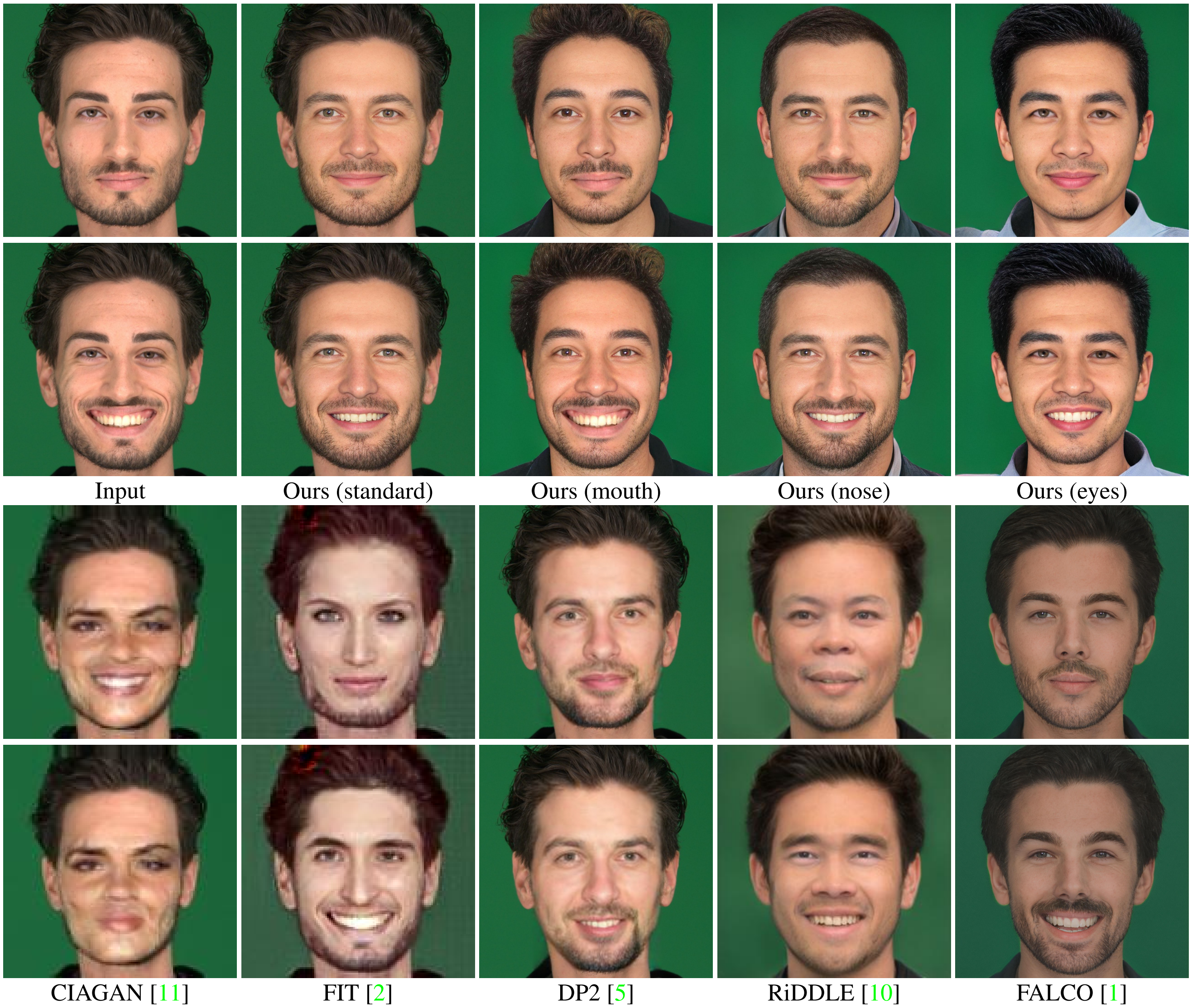
2.5. Downstream utility evaluation

We repeat the downstream utility evaluation presented in our main text on the FFHQ [6] set, and compile the results in Table 2. We achieve the best photorealism and diverse distribution measured by FID, followed by DP2. As for bounding boxes and face detection, we are on par with other state-of-the-art methods, all achieving significantly high performance. The results echo what we present and the conclusions in our main manuscript.

3. Ablation experiments

3.1. Mirroring and projection heads ablation

We perform a simple ablation study over our proposed contrastive mirroring strategy and over our projection heads that are learned on top of each pretrained high-level encoder. We provide the results in Fig. 13. The training contrastive loss for the pose high-level attribute is shown in the top plot. Each curve corresponds to training with our mirroring strategy and projection heads, as well as with the ablation of each. We obtain the best convergence when both components are included. This improved convergence results in improved high-level control, as illustrated visually in the bottom part of the figure. In every row, we sample multiple identities that have a fixed pose latent. Only the



Input Ours (standard) Ours (mouth) Ours (nose) Ours (eyes)

CIAGAN [11] FIT [2] DP2 [5] RiDDLE [10] FALCO [1]

Figure 10. An example pair taken from the SiblingsDB dataset [13] that we anonymize under the *standard* setting and the *clinical* setting preserving the *mouth, nose*, and *eyes* respectively. We compare our results to those of the five benchmarks. Note that in this setting DP2 achieves good identity consistency within the pair. This is because of the highly standardized capture, and the fact that DP2 inpaints the inside of the face conditioned on the outside, which in this setting is almost unchanged.

last row, training with both proposed strategies, consistently achieves the same pose.

3.2. Prior-based blending and correction ablation

The effects of prior-based blending and correction are hard to quantify, since evaluating the photorealism of images (or parts of images) is an open task, tackled only through proxy metrics. We once again use FID [4], accompanied with the downstream utility metrics we compute using face detection models. We perform the ablation study over the postprocessing on the standard anonymization task and provide the results in Tab. 3. We also illustrate the separate and cumulative effects of both steps in Fig. 11, on a sample from our validation set. Face detection models get minimally affected by the changes since only local and low-level features change through the postprocessing and image structures stay mostly similar. We can observe the effects of the postprocessing on FID values on CelebAHQ, but the FID computation with respect to an external dataset (FFHQ, in this example) fails to illustrate the changes in the images. Such small changes are not significant enough to alter the feature distributions with respect to another dataset.

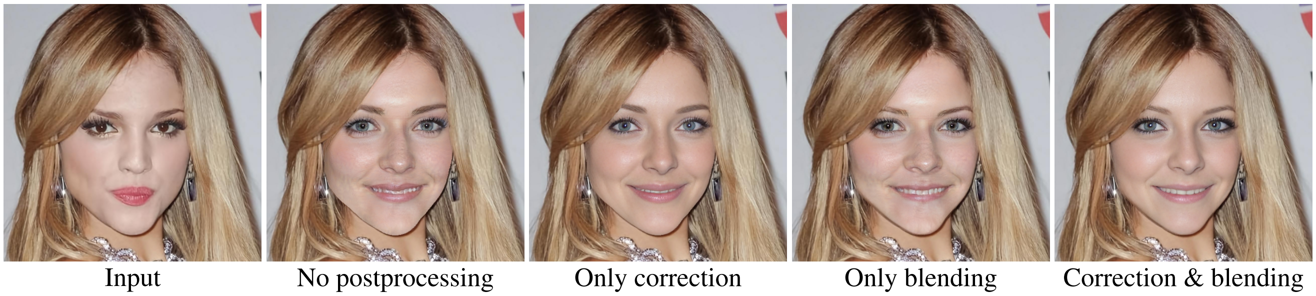


Figure 11. A sample from our ablation experiment over prior-based blending and correction. Prior-based blending fixes the border issues between preserved and non-preserved regions of the image, whereas prior-based correction fixes occasional GAN artifacts and improves texture. Neither step changes the overall structure of the image.

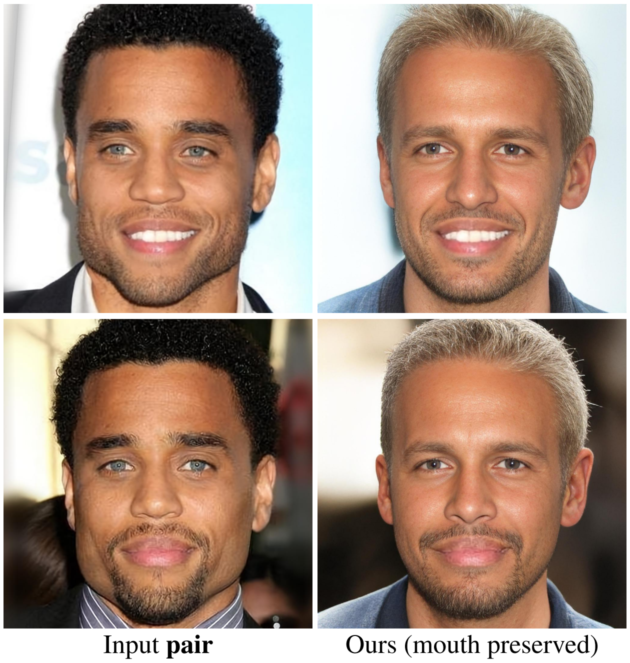


Figure 12. Our *clinical paired-image* anonymization preserving the mouth of the input pair, shown in larger resolution than the main manuscript for better illustration.

Postprocessing method	FID ↓		Bounding box ↑		Face detection ↑	
	FFHQ	CelebAHQ	MTCNN	Dlib	MTCNN	Dlib
No postprocessing	56.92	15.81	0.908	0.955	0.963	0.990
Only correction	57.86	14.03	0.908	0.956	0.966	0.994
Only blending	56.93	13.28	0.907	0.949	0.959	0.990
Correction & blending	57.89	12.49	0.908	0.952	0.962	0.990

Table 3. Effects of prior-based blending and prior-based correction, illustrated through a downstream utility evaluation for photorealism/diversity (FID [4]), bounding box IoU, and face detection rates (MTCNN [14], Dlib [8]) computed over held-out test set from CelebAMaskHQ. Results are comparable with Table 5 from the main manuscript, although there exists a slight difference in the metric for the complete pipeline, accounted by the hardware differences in these two experiments.

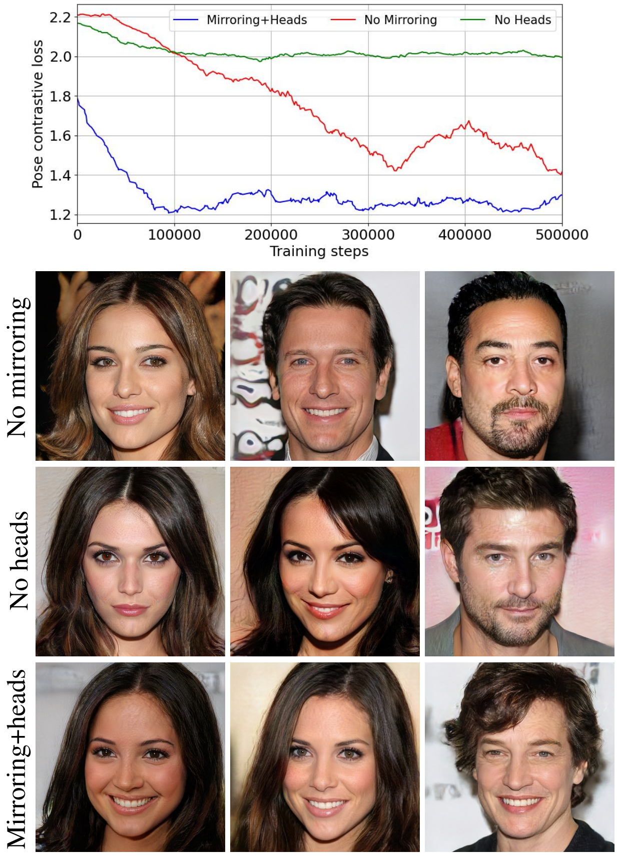


Figure 13. The top plot shows the training contrastive loss for the pose high-level attribute, with our mirroring strategy and projection heads, as well as with the ablation of each component individually. The best convergence is achieved with both of our components. The bottom part illustrates the resulting effect qualitatively. Each row samples multiple identities with a fixed target pose, however, only the last row successfully achieves the same consistent pose across all identities.

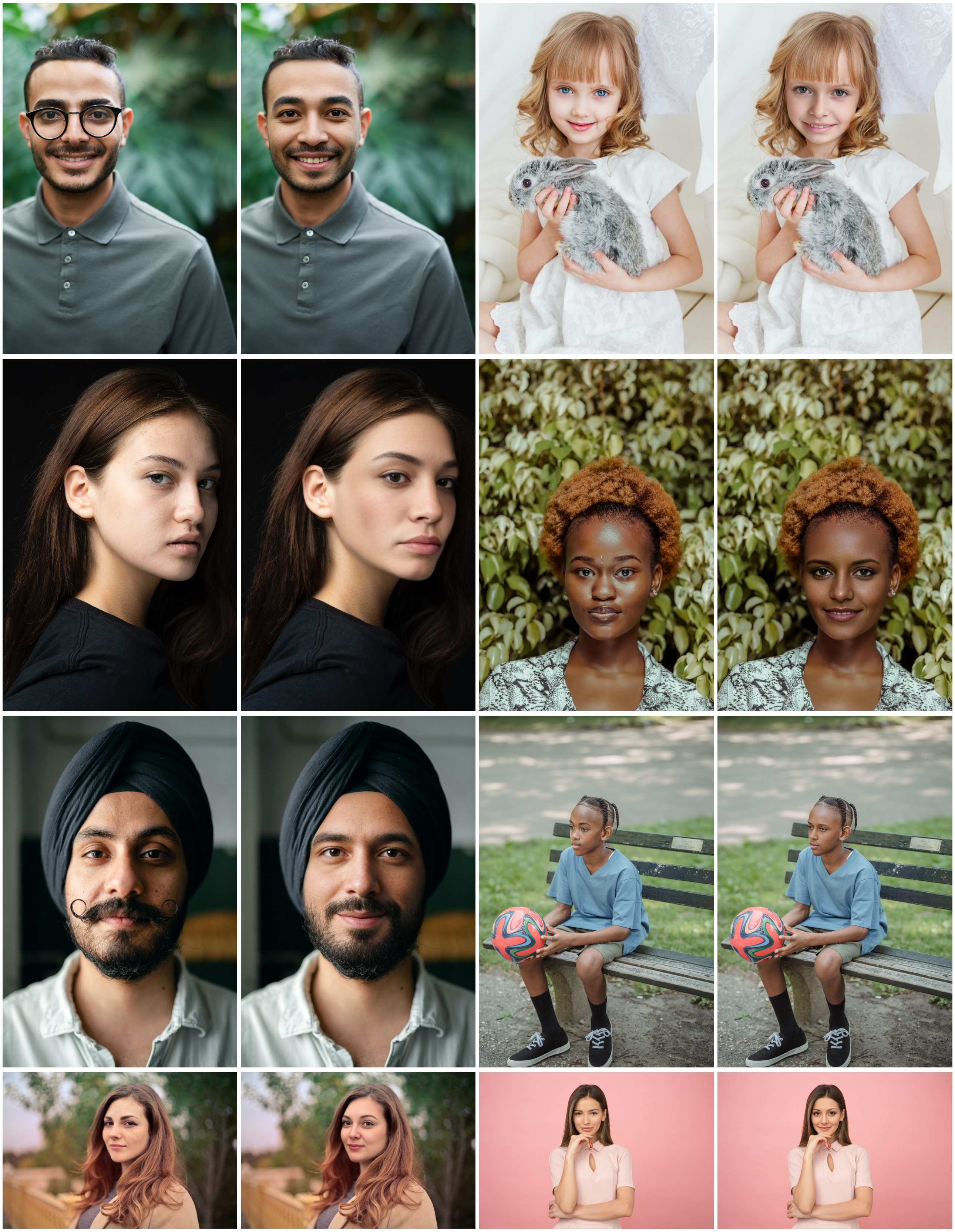


Figure 14. Sample facial anonymization results in full-scene images, all performed using our proposed VerA. VerA works on aligned faces, therefore, we crop the aligned face from the full image as input and replace it by the anonymized face at the same location, following the standard in-place anonymization procedure.

References

[1] Simone Barattin, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. Attribute-preserving face dataset anonymization via latent code optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8001–8010, 2023. 2, 3, 4, 5, 6, 7

[2] Xiuye Gu, Weixin Luo, Michael S Ryoo, and Yong Jae Lee. Password-conditioned anonymization and deanonymization with face identity transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 727–743, 2020. 2, 3, 4, 5, 6, 7

[3] Thorsten Hempel, Ahmed A Abdelrahman, and Ayoub Al-Hamadi. 6D rotation representation for unconstrained head pose estimation. In *IEEE International Conference on Image Processing (ICIP)*, pages 2496–2500, 2022. 2

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 5, 7, 8

[5] Håkon Hukkelås and Frank Lindseth. DeepPrivacy2: Towards realistic full-body anonymization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1329–1338, 2023. 2, 3, 4, 5, 6, 7

[6] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. 2, 4, 5, 6

[7] 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 (CVPR)*, pages 8110–8119, 2020. 1

[8] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 5, 8

[9] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5549–5558, 2020. 2, 3

[10] Dongze Li, Wei Wang, Kang Zhao, Jing Dong, and Tieniu Tan. RiDDLE: Reversible and diversified de-identification with latent encryptor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8093–8102, 2023. 2, 3, 4, 5, 6, 7

[11] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. CIAGAN: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5447–5456, 2020. 2, 3, 4, 5, 6, 7

[12] Yichun Shi, Xiao Yang, Yangyue Wan, and Xiaohui Shen. SemanticStyleGAN: Learning compositional generative priors for controllable image synthesis and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11254–11264, 2022. 1

[13] Tiago F Vieira, Andrea Bottino, Aldo Laurentini, and Matteo De Simone. Detecting siblings in image pairs. *The Visual Computer*, 30:1333–1345, 2014. 5, 7

[14] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 5, 8