Supplementary Material: Custom Structure Preservation in Face Aging

Guillermo Gomez-Trenado[1], Stéphane Lathuilière[2], Pablo Mesejo[1], and Óscar Cordón[1]

[1] DaSCI research institute, DECSAI, University of Granada, Granada, Spain
{guillermogomez,pmesejo,ocordon}@ugr.es
[2] LTCI, Télécom-Paris, Intitute Polytechnique de Paris, Palaiseau, France
stephane.lathuiliere@telecom-paris.fr

In these supplementary materials, we provide additional experiments and visual examples that show the performance of our proposal. First, in Sec. 1, we provide experiments to illustrate the DEX-Face++ age misalignment mentioned in the main paper. Second, in Sec. 2, we further describe our proposed architecture. Third, in Sec. 3, we describe the user study in detail. In Sec. 4, we provide additional qualitative examples, including failure cases and a discussion of examples to complete the ablation study of the main paper. Then, in Sec. 5, we complete our comparison with the State of the Art, including additional results. Finally, in Sec. 6, we list the licenses of the datasets used in our experiments.

1 Age estimation correction

Our evaluation protocol uses Face++ to estimate the person's age in the generated image. However, as mentioned in the main paper, the misalignment of DEX and Face++ classifiers may bias evaluation. In this section, we illustrate how the DEX-Face++ misalignment can bias evaluation.

As in previous approaches [6, 7], DEX is used at training time on the *FFHQ-RR* dataset. Thus, the aging task consists in generating images that match DEX predictions. The DEX-Face++ discrepancy may bias evaluation since an aging method that fails in generating images corresponding to the target age could be favored if the method is biased in the same direction as the Face++ classifier.

To visualize this discrepancy, we plot in Fig 1 the distribution of the DEX-Face++ predictions on the FFHQ-RR dataset. In the case of perfect agreement, all the blue points would be located on the orange identity line. We also report the mean age of each age group according to DEX (red horizontal lines). A vertical dotted line represents the amplitude of the discrepancy. In this case, the discrepancy is especially noticeable in older groups.

Therefore, in our evaluation protocol, we estimate the age of the original images with Face++ and compute the mean for each group. Age MAE is then computed as the distance between the mean group predicted age and the transformed image predicted age.

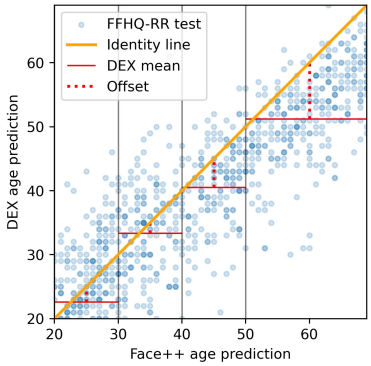

Fig. 1: DEX classifier and Face++ distribution discrepancy by age group on *FFHQ-RR* test set. Color intensity denotes distribution density. The red horizontal lines represent the mean age of each age group according to DEX.

2 Age modulation architecture

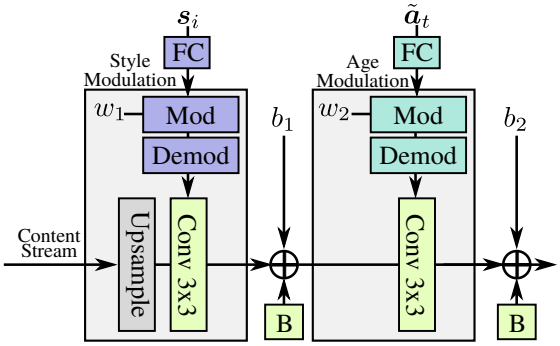

Fig. 2: Illustration of the decoder blocks used in G . B denotes the noise broadcast operation, and FC denotes a fully-connected layer. w_1 and w_2 are two learned scaling parameters, while b_1 and b_2 are learned biases.

In the main paper, we describe our decoder architecture. Its architecture is based on StyleGAN-2 [3], which achieves state-of-the-art performance in unconditional image generation. In addition, we provide several modifications to adapt it to the aging task. These are the use of skip connections, the active manipulation of the skip connection through our CUSP module, and the use of two different inputs in each generator block. The former two are thoroughly discussed in our main work. We now further discuss the latter.

An illustration of our decoder block can be found in Figure 2. Unlike [3], our decoder block takes three inputs: the former block output, the style embedding, and the age embedding. Each decoder block outputs an image twice the size of its input and is composed of two consecutive sub-blocks: the *style sub-block* and the *age block*. In the *style sub-block*, the input is upscaled through bilinear interpolation. Then the upscaled input is transformed through weight demodulation (w_1) based on a linear combination of the style embedding (s_i). In the second sub-block, the age embedding $\tilde{a}_t$ and w_2 are used for transforming the former sub-block output. Both s_i and $\tilde{a}_t$ are shared by every block. After each step, 0-centered random noise B is added to the output.

3 User study

We now provide some details regarding the user study reported in the main. Each test consisted of 48 random questions on four different topics. In total, 72 users were evaluated. Similarly to [4] approach this is the description prompted to the users.

> In this study, you will be presented with several sets of images to choose from. We will compare several AI solutions to transform a person's age in an image, similar to widely known apps like FaceApp. There are four kinds of questions, you'll have to click on your chosen image, there are no correct answers:
>
> 1. **Age accuracy:** From the images displayed, which one better depicts a person from the *target age group*? An actual person's picture (not shown) has been transformed to a target age with different mechanisms. We want to know which one you think is more accurate.
> 2. **Identity preservation:** From the images displayed, which one better transforms the shown original picture to the target age group while *reasonably maintaining the person's identity*? You'll have to judge which result seems more reasonable, attending to age transformation and identity preservation.
> 3. **Overall better:** From the images displayed, which one is *overall better* transforming the age of the person depicted in the picture? Which one do you prefer? Which image seems more pleasing?
> 4. **Whole age progression:** From the different shown *age progressions*, which seems *more natural and reasonable*?
>
> In case of doubt, choose the image you subjectively prefer.

From FFHQ-RR, 50 images were selected for each group (20-29, 30-39,40-49,50-69) and transformed to target ages 25, 35, 45, 60 with each comparing method (HRFAE [6], LATS [4], and ours), resulting in 200 original images and 2400 transformed images. Every image from our method was obtained with CP configuration $(\sigma_m, \sigma_g) = (8, 1.8)$. Age translations were done from 20-29 and 30-39 to 50-69 (young to old) and from 40-49 and 50-69 to 20-29 (old to young).

In Question-Kind 1 (QK 1) and QK 3, three randomly ordered transformed images were presented next to a target age group. In QK 2, the original image is included. Finally, in QK 4, besides the original image, four images showing age progression (25, 35, 45, and 60) are presented for each method.

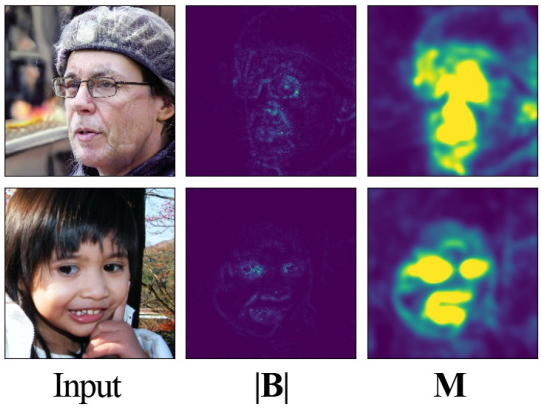

Fig. 3: Example outputs of the CUSP module. From left to right: 1) Input image, 2) Matrix $|\mathbf{B}|$, the absolute value of the guided backpropagation output averaged over the RGB dimension, 3) Mask $\mathbf{M}$ predicted by the CUSP module.

4 Ablation Study

4.1 CUSP processing

We now provide some visualizations that motivate the proposed computation for the mask $\mathbf{M}$. Figure 3 shows the output $\mathbf{B}$ of the Guided backpropagation algorithm for two input images (2nd column). We see that $\mathbf{B}$ is very sparse. Therefore, we apply blur before normalization and clipping to enlarge the activated regions. In this way, we obtain the mask in the last column. We see that the high values of the masks are primarily located in the eye and mouth regions, while the background is associated with very low values. This visualization shows that our CUSP module can act only on the relevant regions in the foreground.

Furthermore, we compared our CUSP module with supervised alternatives such as segmentation-based masking [2, 4]. Even though our predicted mask is not always accurate, it has several advantages: *(i)* It rules out the need for extra supervision (*e.g.*, landmarks); *(ii)* CUSP with the Custom Preservation (CP) setting targets only age-specific regions while face segmentation uniformly blurs the whole face area; *(iii)* When we evaluated a segmentation model (*BiSeNet* trained on *CelebAMask-HQ*) as the CUSP mask, it showed that the segmentation-based mask introduces new artifacts (see left ear on Fig. 4) probably due to mask inaccuracies. Regarding CUSP, Low Preservation (LP) setting losses background details, but CP manages to preserve the background despite the inaccurate mask.

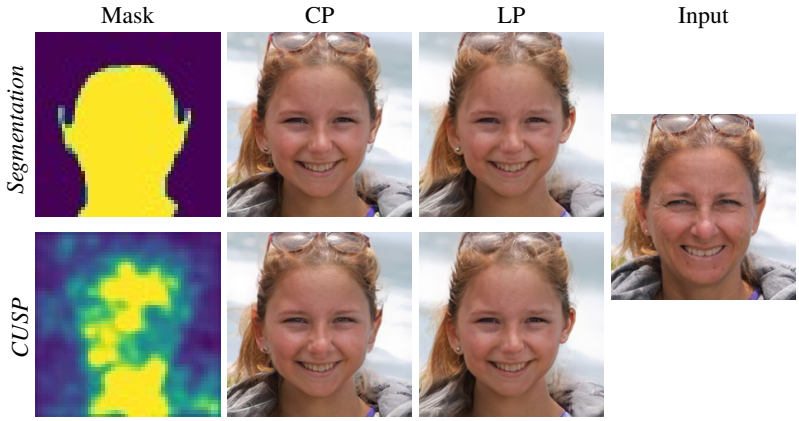

Fig. 4: Qualitative evaluation of segmentation-based masking.

4.2 Architecture ablation

Figure 5 introduces the corresponding images for the qualitative ablation study made in our main work for the architecture design. It can be observed that the final architecture (*Full*) retains identity, details, and gender better than the alternatives, even in challenging examples such as the second image.

On the other hand, in Figure 6, the images corresponding to the masking strategy ablation are shown. Even though the differences are subtle, the class-independent approach presents fewer artifacts while preserving details and identity better.

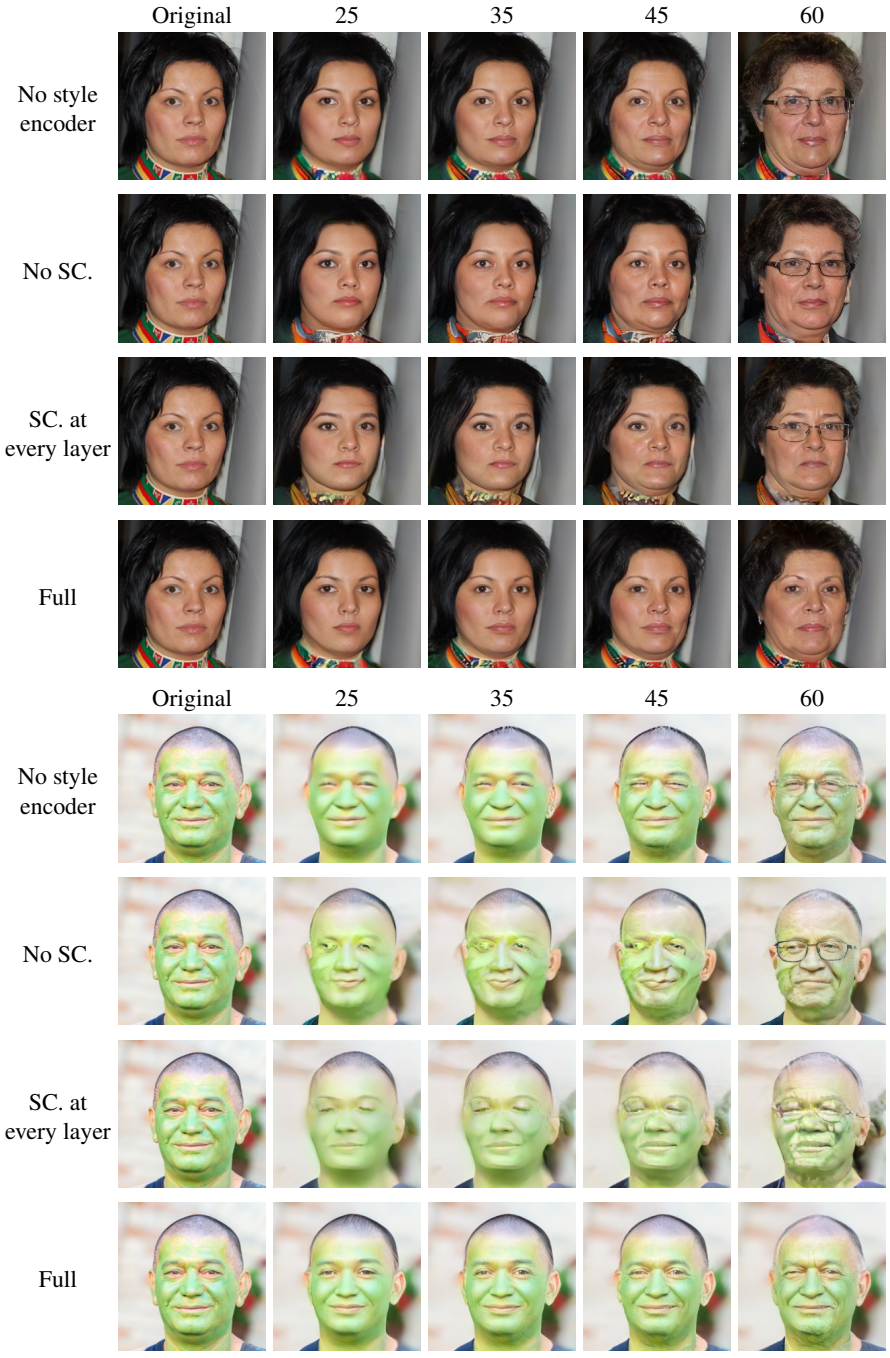

Fig. 5: Ablation study: impact of the skip connections (SC.) and the style encoder

Fig. 6: Ablation study: impact of the masking strategy used in CUSP.

4.3 Additional quantitative results

In this section, we quantitatively study the impact of the blurring approach. We compare four approaches: the three masking configurations employed in the experiments of the main paper (HP, CP, and LP) and a *Global blur* approach that uniformly blurs all images during training. This *Global blurring* approach is equivalent to setting σ_m to 0 and σ_g to 9. Qualitative results are reported in Table 1.

Table 1: Blurring approach ablation on the *FFHQ-RR* test set. CUSP HP (High preservation), CP (Custom preservation), and LP (Low preservation) are run with $(\sigma_m, \sigma_g) = (0, 0)$, $(\sigma_m, \sigma_g) = (8, 1.8)$ and $(\sigma_m, \sigma_g) = (8, 4.5)$ respectively. *Global blur* is trained and run with $(\sigma_m, \sigma_g) = (0, 9)$.

	Reconstruction	*Age translation*	
	LPIPS ($\times$10)	**Age MAE**	**Mean FID**
Global blur	1.56	**5.84**	109.03
CUSP - LP	1.09	6.07	**104.44**
CUSP - CP	0.78	6.29	104.63
CUSP - HP	**0.71**	9.05	106.78

It shows that the two models with lower structure preservation (*Global blur* and LP) obtain the best age translation scores, while their high reconstruction error (*i.e.*, LPIPS) shows that the details of the images are not preserved. On the contrary, CUSP-HP achieves better reconstruction at the cost of worse age translation scores. Our CUSP model with Custom blur parameters leads to a satisfying trade-off between reconstruction and age translation.

These experiments again justify the usefulness of letting the user the possibility to choose its own trade-off at inference time, as well as the use of different values for σ_m and σ_g.

4.4 Additional qualitative results

We now qualitatively evaluate the three masking configurations employed in the experiments of the main paper (HP, CP, and LP). Results on the *FFHQ-RR* dataset are shown in Figs. 8 and 7 for two different settings, old to young and young to old respectively. Similar to the main paper, these results show that the high structure preservation variant preserves the face shape and hair growth. Meanwhile, the lower structure preservation allows stronger modifications of the face. We can see that the intermediate model with custom preservation achieves a satisfying trade-off.

Finally, both CP and LP are evaluated on six different target ages and two CUSP configurations in Figs. 9 and 10 on the *FFHQ-RR* dataset. Again, we see that lower

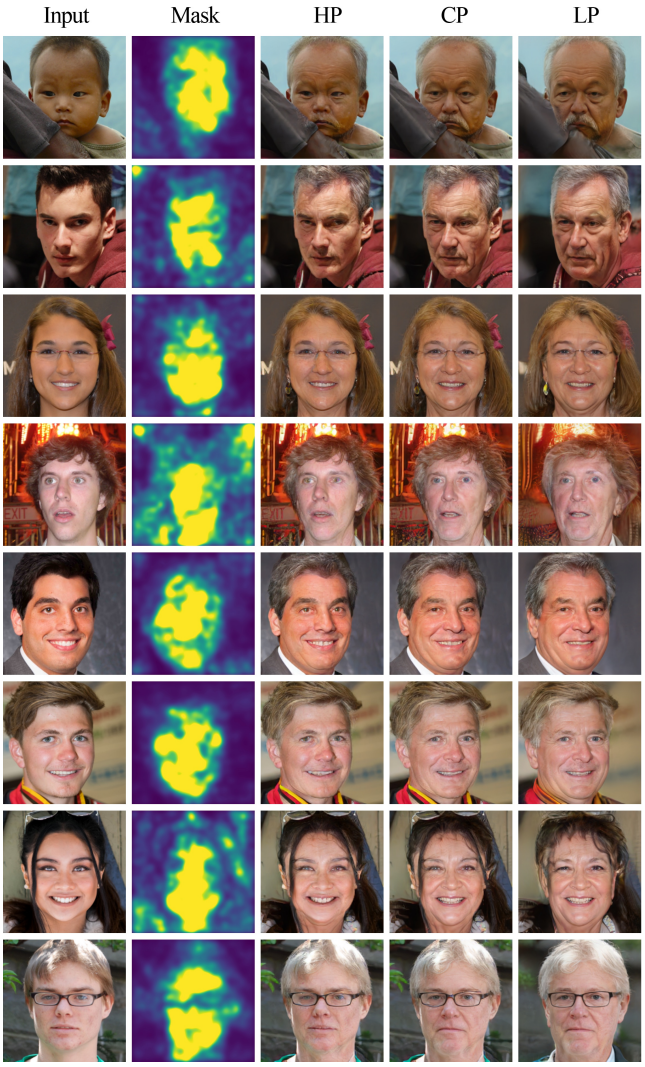

Fig. 7: Qualitative study of the impact of the kernel value in CUSP at test time on young (lower than 30) to old (60). We compare images obtained with High, Low, and Custom structure preservation (referred to as LP, HP, and CP). HP:$(\sigma_m, \sigma_g) = (0, 0)$; CP:$(\sigma_m, \sigma_g) = (9, 0)$; HP:$(\sigma_m, \sigma_g) = (9, 9)$. The second column shows the mask estimated by our CUSP module with a color scale from blue (for 0) to yellow (for 1).

preservation (*i.e.*, LP) applies strong modifications to the images to match the target age at the cost of lower preservation of identity (*e.g.*, see hairs in the first row of 10)

4.5 Failure cases

In Fig. 11, we show some failure cases. These images present noticeable artifacts. We observe that these artifacts appear mainly in the presence of white beards. The CUSP mask does not entirely select beards (see the second row). This behavior might indicate an inherited bias from IMDb-Wiki, where white beards can be rare in celebrities' pictures.

In these images, we also see some saturation artifacts similar to those described in [3]. Even though we employ *Weight demodulation* [3] that is aimed at solving this issue, we observe that some artifacts remain.

5 Comparison with State-of-the-Art

We now report additional qualitative comparison with State-of-the-Art. We report separate comparisons with HRFAE and LATS as in the main paper.

5.1 Comparison with HRFAE

In Fig. 12, we show some additional comparisons with HRFAE on the *FFHQ-RR* dataset using the CUSP CP setting. These results are in line with the results reported in the main paper. In addition, we observe that our approach is able to apply more substantial modifications to the image to better match the target age.

We complete this comparison by showing the results obtained with our method on the same examples previously used in [6] and using their qualitative evaluation. We generate images varying the target age from 20 to 69. Our results are smooth, and we observe that our proposed method is able to apply more profound changes in the case of extreme ages.

5.2 Comparison with LATS

In Fig. 14, we show a complementary qualitative comparison with LATS on several images. Similar to the results in the main paper, we observe that our method is on par with LATS while having the numerous advantages detailed in the main paper.

Pending comparison

Posteriorly to the submission of this work, [1] was published. The authors propose a new method for age editing built upon a collection of pretrained networks and custom-trained modules (methodology described in our main paper). Even though they inherit the same flaws pretrained StyleGAN2 and pSp [5] have (*i.e.*, blurry backgrounds and low structural and identity preservation), they achieve promising deep structural age-related transformations. For this, comparing both methods in the future should be interesting.

6 Datasets licenses

Both CelebA-HQ and FFHQ are publicly available and widely used datasets. CelebA-HQ is openly available for its use in research but has some rights reserved. FFHQ is made available under the Creative Commons BY-NC-SA 4.0 license. Thus every derivative (e.i., FFHQ-RR and FFHQ-LS) have the same license.

References

1. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Only a matter of style: Age transformation using a style-based regression model. ACM Trans. Graph. **40**(4) (2021)
2. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
3. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: IEEE/CVF CVPR (2020)
4. Or-El, R., Sengupta, S., Fried, O., Shechtman, E., Kemelmacher-Shlizerman, I.: Lifespan age transformation synthesis. In: IEEE/CVF ECCV (2020)
5. 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: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
6. Yao, X., Puy, G., Newson, A., Gousseau, Y., Hellier, P.: High resolution face age editing. In: IEEE ICPR (2021)
7. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: IEEE/CVF CVPR (2017)

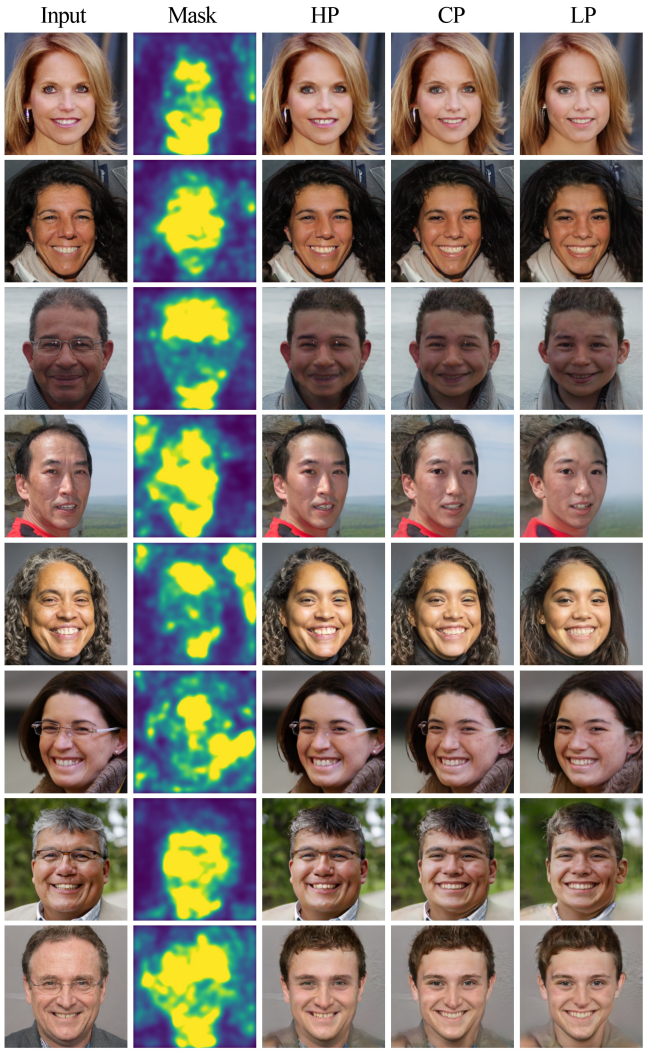

Fig. 8: Qualitative study of the impact of the kernel value in CUSP at test time on old (between 50 and 60) to young (25). We compare images obtained with High, Low, and Custom structure preservation (referred to as LP, HP, and CP). HP:$(\sigma_m, \sigma_g) = (0, 0)$; CP:$(\sigma_m, \sigma_g) = (9, 0)$; HP:$(\sigma_m, \sigma_g) = (9, 9)$. The second column shows the mask estimated by our CUSP module with a color scale from blue (for 0) to yellow (for 1).

Fig. 9: Qualitative study of the impact of kernel values in CUSP at test time on the *FFHQ-RR* test set. For each rows pair, the first corresponds to Custom structure preservation or CP $(\sigma_m, \sigma_g) = (0.9, 7.2)$, the second to Low preservation, LP $(\sigma_m, \sigma_g) = (8.6, 7.2)$.

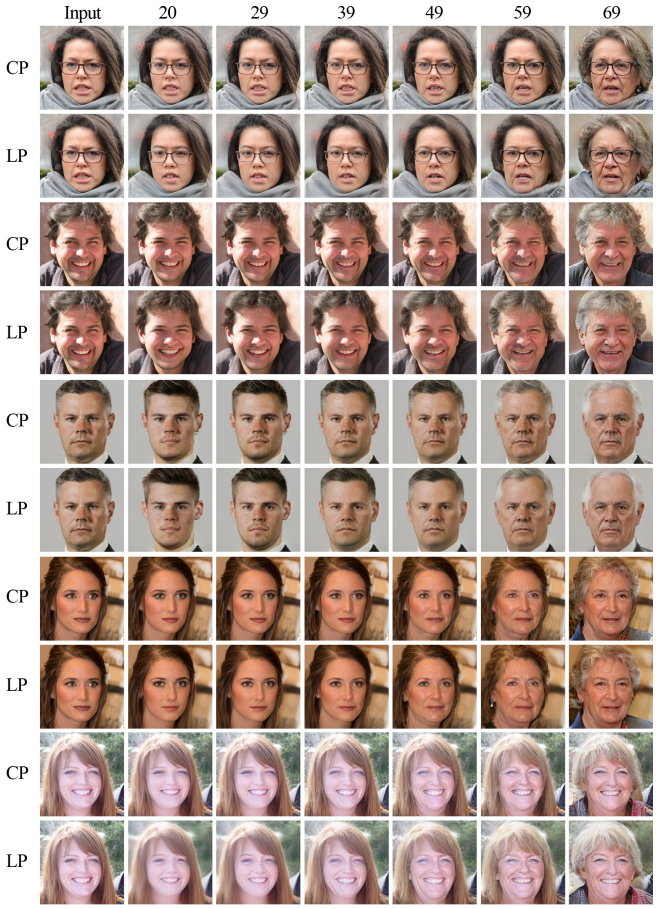

Fig. 10: Qualitative study of the impact of the kernel value in CUSP at test time on the *FFHQ-RR* test set. For each rows pair, the first corresponds to Custom structure preservation $(\sigma_m, \sigma_g) = (0.9, 7.2)$, the second to Low preservation $(\sigma_m, \sigma_g) = (8.6, 7.2)$.

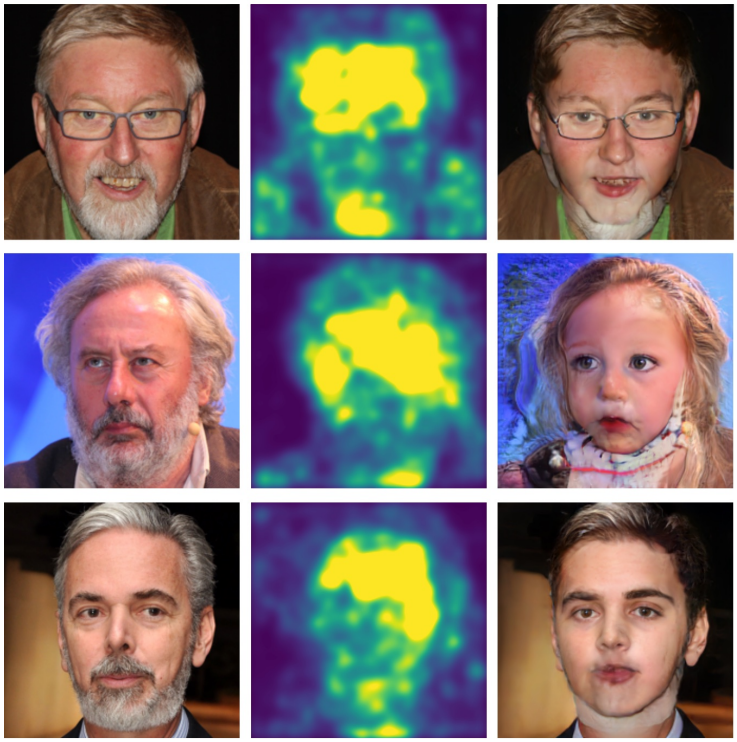

Fig. 11: From left to right: Input image, **M** Mask, and target age 25. The resulting images are obtained with the CUSP CP setting $(\sigma_m, \sigma_g) = (8, 1.8)$ on *FFHQ-RR*.

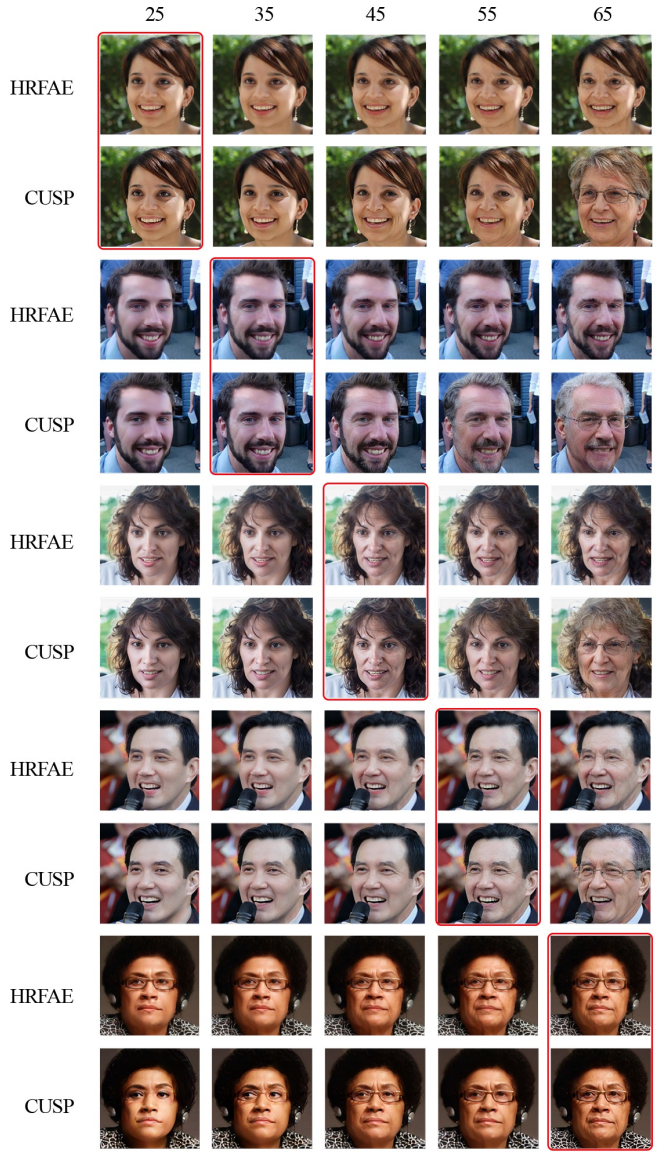

Fig. 12: Qualitative comparison with HRFAE for different age targets on *FFHQ-RR*. The used setting is CUSP CP $(\sigma_m, \sigma_g) = (8, 1.8)$. The images corresponding to the target ages are highlighted with red frames.

Fig. 13: HRFAE comparison on *FFHQ-RR* for smooth progression. The target age goes uniformly from 20 to 69.

Fig. 14: Qualitative comparison with LATS on *FFHQ-LS* test set for different age targets.