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

Abstract. In this work, we propose a novel architecture for face age editing that can produce structural modifications while maintaining relevant details present in the original image. We disentangle the style and content of the input image and propose a new decoder network that adopts a style-based strategy to combine the style and content representations of the input image while conditioning the output on the target age. We go beyond existing aging methods allowing users to adjust the degree of structure preservation in the input image during inference. To this purpose, we introduce a masking mechanism, the CUstom Structure Preservation module, that distinguishes relevant regions in the input image from those that should be discarded. CUSP requires no additional supervision. Finally, our quantitative and qualitative analysis which include a user study, show that our method outperforms prior art and demonstrates the effectiveness of our strategy regarding image editing and adjustable structure preservation.

Keywords: Face aging, Image editing, Style-base architecture

1 Introduction

Face age editing [7, 17, 39], or aging, consists in automatically modifying an input face image to alter the age of the depicted person while preserving identity. Over the last few years, this problem has attracted a growing interest because of its numerous applications. In particular, it is used in the movie production industry to edit actors' faces or in forensic facial approximation to reconstruct the faces of missing people. The advances in deep learning methods unlock the development of fully automatic edition algorithms that avoid hours of makeup and post-production retouching.

Recent deep learning approaches adopt an encoder-decoder architecture [3, 8, 23, 26, 39, 40, 43, 45]. The image is encoded in a latent space that can be modified depending on the target age and fed to a decoder that generates the output image. The overall network is usually trained using a combination of losses that assess image quality, identity preservation, and age matching. However, despite the success of all these approaches, face editing remains challenging, and current methods usually fail when

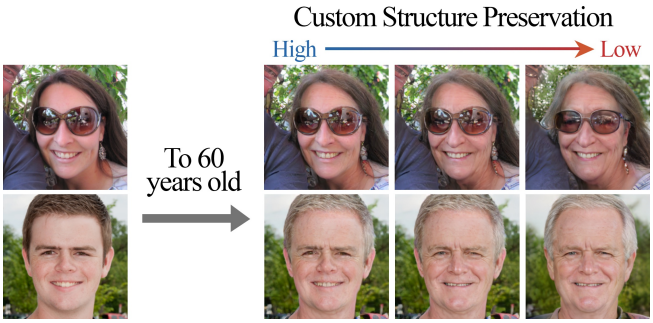

Fig. 1: The user can choose the degree of structure preservation at inference time. Facial morphology transformations are more profound as we move to the right (lower structure preservation).

faced with sizeable differences between the age of the person displayed in the input image and the target age. Indeed, most approaches [3, 8, 39, 40, 43, 45] only superficially modify the skin's texture while the face's shape is kept unchanged. These approaches fail with significant age gaps since face shape can change significantly during a lifetime. Few methods try to go beyond some limited age gaps, but they either consider only a tightly cropped face region [17, 39] or require specific pre-processing involving an image segmentation step [26].

This work proposes a novel framework that allows profound structural changes in facial transformations. This framework achieves a realistic image transformation with age gaps that imply changes in head shape or hair growth. In addition, we argue that the face editing task is an ill-posed problem because every person gets older in a different and non-deterministic way: some people drastically change, while others are easily recognizable in old photographs. In this sense, we propose a methodology that allows the user to adjust, at inference time, the degree of structure preservation. Thus, the user can provide an image and obtain different transformations where the structure (*i.e.*, face shape or hair growth) is preserved at different levels. Fig. 1 shows some qualitative results obtained with our method. Furthermore, the user can choose different degrees of structure preservation: with high preservation, the model only changes the texture, while with lower preservation, the shape of the face is also modified.

The contributions of this paper can be summarized as follows:

- We propose a novel architecture for face age editing that can produce structural modifications in the input image while maintaining relevant details present in the original image. We take advantage of recent advances in image-to-image (I2I) translation [10, 20] and unconditional image generation [12] to design our architecture. We disentangle the style and content of the input image, and we propose a new decoder network that adopts a style-based strategy to combine the style and content representations of the input image while conditioning the output on the target age.
- We go beyond existing aging methods allowing the user to adjust the degree of structure preservation in the input image at inference time. To this aim, we intro-

duce a masking mechanism, through a so-called CUstom Structure Preservation (CUSP) module, that identifies the relevant regions in the input image that should be preserved and those where details are irrelevant to the task. Importantly, our mechanism for adjustable structural preservation does not require additional training supervision.
- Experimentally, we show that our method outperforms existing approaches in three publicly available high-resolution datasets and demonstrate the effectiveness of our mechanism for adjusting structure preservation.[3]

2 Related Work

Most recent approaches for **face aging** adopt a similar strategy based on an encoder-decoder architecture [3, 8, 23, 26, 39, 40, 43, 45]. In these methods, the input image is projected onto a latent space where content is manipulated before decoding the output image. Some methods [40, 2] add an identity term to the total loss to better ensure the preservation of the identity during the translation process. These methods principally differ in the choice of the network architecture and the manner the latent representation is manipulated. For instance, Wang *et al*. [39] introduce a recurrent neural network to iteratively alter the image, while in [43], the latent image representation is modified using a simple affine transformation. Re-AgingGAN [23] employs an age modulator that outputs transformations that are applied then to the decoder, and Or-El *et al*. [26] adopt a multi-domain translation formulation, showing that segmentation information can be leveraged to improve aging. In our work, we adopt an encoder-decoder framework similar to [8, 43]. However, our approach goes beyond existing methods that generate a single image for a given image-target age pair. Indeed, we offer the user the possibility to adjust the degree of structure preservation during translation, and, in this way, we can output a set of plausible resulting facial images.

Our method also leverages recent advances from the **I2I translation** research area. I2I translation consists in learning a mapping between two visual domains. In the pioneering work of Isola *et al*. [11], an encoder-decoder network is trained using a dataset composed of image pairs from the two domains. Later, many works addressed I2I translation in an unpaired setting, assuming two independent sets of images of each domain [6, 21, 46]. These works, of which cycleGAN [46] is a paradigmatic example, mainly focus on introducing regularization mechanisms when training the I2I translation models. Another research direction is designing more advanced architectures to improve image quality or obtain several possible outputs for a given input [10, 20, 47]. Disentangling style and content information has led to both higher image quality and diversity [10, 28]. We adopt a similar strategy in order to allow custom structure preservation. Thanks to this strategy, our CUPS module can act on the spatial information passing through the content branch while preserving style information.

[3] Code and pretrained models are available at `https://github.com/guillermogotre/CUSP`.

Style-based architectures recently attracted much attention for the problem of unconditional image generation. In particular, StyleGAN2 [12] is now used in many face manipulation tasks [30, 42]. In the case of face aging, [2] uses a pretrained StyleGAN2 model [12] equipped with a pSp encoder [30], and an age classifier [32] to tailor an age editing model with unlabeled data. In StyleGAN2, a network maps a Gaussian latent space onto style vectors; these vectors are later combined via a convolutional network to produce the output image. Finally, the synthesis network aggregates the style vectors through modulation operations. We take inspiration from the StyleGAN2 generator to design a novel decoder that combines the input style and the target age with the content representation via weight demodulation.

Regarding the more general **image editing** problem, our method shares similarities with several approaches employing masking mechanisms or attention maps to preserve relevant parts in the input image [1, 18, 29, 38]. For instance, mask consistency is employed in [18] to improve multi-domain translations. As in our approach, masks are estimated using the guided backpropagation (GB) algorithm [36]. In the case of facial images, a mask is employed in GANimation [29] to different regions that should be preserved and those that should be modified to change the facial expression. In GANimation, masks are predicted by the main network, while we employ an auxiliary network and GB [36] to obtain the mask.

3 Proposed Method

In this work, we address the face age editing problem. Therefore, our goal is to train a network able to transform an input image $\mathbf{X}$, such that the person depicted looks like being of the target age a_t. At training time, we assume that we have at our disposal a dataset composed of I face images of resolution $H \times W$, such that $\mathbf{X}_i \in \mathbb{R}^{H \times W \times 3}, i = 1, ..., I$ with their corresponding age label $a_i \in \{1, ..N\}$. Note that the age labels are automatically obtained using a pre-trained age classifier. Similar to previous approaches [43, 45], we employ the DEX classifier [33].

One of the main difficulties lies in modifying the relevant details in the input image while preserving non-age-related regions. To this aim, we introduce a style-based architecture detailed in Sec.3.1. In contrast to previous works, the CUSP module allows the user to indicate the desired level of structure preservation through two parameters: $\sigma_m > 0$ and $\sigma_g > 0$. These parameters act locally and globally, respectively, as detailed later. The proposed CUSP module is described in Sec. 3.2. Finally, we present the whole training procedure in Sec. 3.3.

3.1 Style-based Encoder-decoder

As illustrated in Fig. 2, our architecture employs five different networks: (1) A style encoder E_s extracts a style representation $\boldsymbol{s}_i$ of the input image $\mathbf{X}_i$. E_s discards any spatial information via global-average-pooling at the last layer. The use of a style encoder allows global information to be used at any location in the decoder. (2) A content

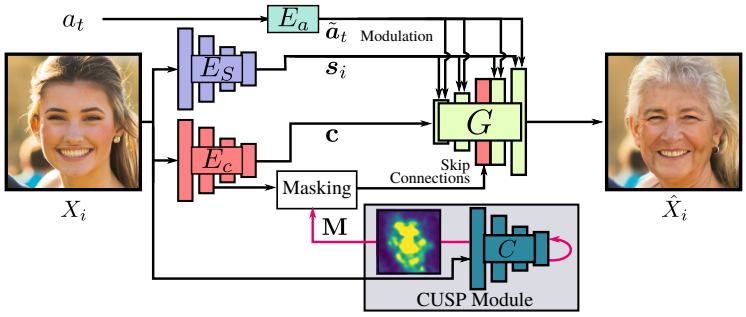

Fig. 2: Illustration of the proposed approach. A style encoder E_s extracts a style representation of the input image $\mathbf{X}_i$. A content encoder E_c encodes spatial information. Target age a_t is embedded using a multi-layer perceptron E_a. Our generator G outputs the image $\hat{\mathbf{X}}_i$ by combining the input style and content representations conditioned on the target age. Our CUSP module predicts a blurring mask $\mathbf{M}$ applied to the skip connections to allow the user to choose a CUstom level of Structure Preservation.

encoder E_c outputs a tensor $\mathbf{c}$ describing the content of the input image. Contrary to E_s, the content encoder preserves spatial and local information. In our case, the use of separated style and content encoders is justified by the fact that our CUSP module should not affect the image style $\boldsymbol{s}_i$ but only the structure of the image. (3) An 8-layers fully connected network, E_a, embeds the target age a_t: $\tilde{\boldsymbol{a}}_t = E_a(a_t)$. (4) An image generator G estimates the output image $\hat{\mathbf{X}}_i$ by combining the style and content representations with the target age embedding $\tilde{\boldsymbol{a}}_t$. (5) Finally, our CUSP module allows the user to choose the level of structure preservation. This module predicts a mask $\mathbf{M}$ used to act on the skip connections between the content encoder and the decoder. More precisely, we blur the regions indicated by the mask $\mathbf{M}$ to propagate only the non-age-related structural information to the decoder.

Our image generator G is designed to combine the outputs of the style and content encoders with the target age embedding. Its architecture is inspired by StyleGAN2 [16], which achieves state-of-the-art performance in unconditional image generation. However, we provide several modifications to tailor the architecture to the aging task. G comprises a sequence of elementary blocks (see *Supplementary Materials* for illustration). Differently from StyleGAN2, each block takes three inputs: the former block output feature map, the style encoding $\boldsymbol{s}_i$, and the class embedding $\tilde{\boldsymbol{a}}_t$. Each block is composed of two sub-blocks. In the first one, we use the style vector $\boldsymbol{s}_t$ to modulate the convolution operations as in [16]. In the second one, the age embedding is used for modulation. Up-sampling is applied to the input of the first sub-block. Similarly to [16], random noise is summed to the feature maps between each sub-block, while scaling and bias parameters (*i.e.*, w and b) are learned for each sub-block.

Note that all blocks are combined following the *input skips* architecture of StyleGAN2, where a layer named *tRGB* is introduced. Such layer predicts intermediate images at every resolution scaled and added to generate the final image. tRGB is also

conditioned on the age embedding. A single skip connection is introduced before the last block, contrarily to U-Net [31] that includes them in every layer.

3.2 CUSP Module

Skip connections (SC) [31] are efficient tools to provide high-frequency information to the decoder allowing accurate reconstruction [11]. High frequencies carry accurate spatial information that favors pixel-to-pixel alignment between inputs and outputs, as, for instance, needed in segmentation. However, previous works [35] show they are not suited for tasks where the input and output images are not pixel-to-pixel aligned. For example, input and output images are aligned when the age gap is small in the aging task. However, this assumption does not hold in every image region with significant gaps. This misalignment is particularly predominant in areas other than the background since facial morphology or hairstyle may change.

Therefore, we propose to control the amount of structural information that flows through the SC. This control is obtained by blurring the feature maps going through them. Nevertheless, every region should not be treated in the same way. For instance, depending on the task, the user may prefer to preserve the background while blurring the foreground to loosen conditioning on the input image in this region. Therefore, we propose a specific mechanism to identify relevant image regions for the translation.

Mask Estimation. We employ an additional classification network C, pretrained to recognize the age of the person depicted on an image. We use the DEX classifier [33] again. Since DEX is pretrained on 224×224, the input image is rescaled to this resolution. Then, we apply the GB algorithm [36] to obtain a tensor $\mathbf{B} \in \mathbb{R}^{224\times224\times3}$, where locations with higher norm correspond to regions predominantly used by the classifier. In other words, $\mathbf{B}$ pinpoints relevant regions for the age classification task. GB points out the key areas to recognize the age and should, therefore, be modified by the aging network. Importantly, GB is usually used to visualize the regions that influence one specific network output (*i.e.*, one specific class) [36]. In our case, we apply GB to the sum of the classification layer before softmax normalization to obtain class-independent masks. We select GB [36] over other approaches [34, 25, 37] since it is a fast, simple, and strongly supported method for visualization.

We need to transform $\mathbf{B}$ to obtain a mask $\mathbf{M} \in [0,1]^{224\times224}$. We proceed in several steps. First, we average $\mathbf{B}$ over the RGB channels, take the absolute value, and apply Gaussian blur to get smoother maps. In this way, we obtain a tensor $\tilde{\mathbf{B}} \in {\mathbb{R}_{>0}}^{224\times224}$ that indicates relevant regions. To obtain values in $[0,1]$, we need to normalize $\tilde{\mathbf{B}}$. Our preliminary experiments showed that after normalizing by twice the variance σ of $\tilde{\mathbf{B}}$ (over the locations), relevant areas for the aging task are close to 1 or above. We apply clipping to bring all those important regions to 1. Formally the mask values are computed as follows:

$$\mathbf{M} = \min\left(\frac{\tilde{\mathbf{B}}}{2\times\sigma}, 1\right) \tag{1}$$

where *min* denotes the element-wise minimum. Next, we detail how this mask is employed in our encoder-decoder architecture.

Skip connection blurring. Assuming a feature map $\mathbf{F}_c \in \mathbb{R}^{H' \times W' \times C}$ provided by the content encoder E_c, we resize $\mathbf{M}$ to the dimension of $\mathbf{F}_c$ obtaining a mask $\tilde{\mathbf{M}} \in [0,1]^{H' \times W'}$. We then blur $\mathbf{F}_c$ using two different Gaussian kernels with variance $\sigma_m > 0$ and $\sigma_g > 0$. The variance σ_m is applied in the region indicated by $\mathbf{M}$, while σ_g is used over the whole feature map. The motivation for this choice is that the user can choose to alter structure preservation locally, globally, or both. At training time, σ_m and σ_g are sampled randomly to force the generator G to perform well for any blur parameter. At test time, both values might be provided by the user. Formally, the blurred feature map is computed as follows:

$$\tilde{\mathbf{F}}_c = \tilde{\mathbf{M}} \circ (\mathbf{F}_c * \mathbf{k}_m) + (1 - \tilde{\mathbf{M}}) \circ (\mathbf{F}_c * \mathbf{k}_g) \tag{2}$$

where $*$ denotes the convolution operation, $\circ$ is the Hadamard product, and $\mathbf{k}_m$ and $\mathbf{k}_g$ are the Gaussian kernels of variances σ_m and σ_g.

3.3 Overall Training Procedure

Training facial age editing models is particularly challenging since paired images are unavailable. Therefore, similarly to [23, 26, 43], our training strategy is either focused on reconstruction (when the target age matches the input age) or I2I translation (when the target age is different). Also, similar to [23, 26, 43], training is performed using a set of complementary losses described below.

Reconstruction loss ($\mathcal{L}_r$). When the target age a_t is equal to the image age a_i, we expect to reconstruct the input image. We, therefore, adopt an L1 reconstruction loss:

$$\mathcal{L}_r = \|T(\mathbf{X}_i, a_i) - \mathbf{X}_i\|_1 \tag{3}$$

where T denotes the whole aging network, which output is the scaled addition of every *tRGB* block.

Age fidelity losses ($\mathcal{L}_D, \mathcal{L}_C$). Following [5], we use a conditional discriminator D to asses that generated images correspond to the target age a_t. More precisely, we employ the discriminator architecture of StyleGAN2 equipped with a multiclass prediction head, together with the training loss $\mathcal{L}_D$ defined in [24].

We employ a loss $\mathcal{L}_C$ that assesses age matching using the same pretrained classifier C used in the CUSP module to complement the adversarial loss. Furthermore, $\mathcal{L}_C$ is implemented using the Mean-Variance loss [27], a classification loss tailored for age estimation.

Cycle-Consistency loss ($\mathcal{L}_{cy}$). Following [46], we adopt a cycle consistency $\mathcal{L}_{cy}$ to force the network to preserve details that are not specific to the age (*e.g.*, , background or face identity). $\mathcal{L}_{cy}$ is given by:

$$\mathcal{L}_{cy} = \|\mathbf{X}_i - T(T(\mathbf{X}_i, a_t), a_i)\|_1 \tag{4}$$

Full objective. Finally, the total cost function can be written

$$\min_{M} \max_{D} \lambda_r \mathcal{L}_r + \lambda_C \mathcal{L}_C + \lambda_D \mathcal{L}_D + \lambda_{cy} \mathcal{L}_{cy} \tag{5}$$

where $\lambda_r, \lambda_C, \lambda_D$, and λ_{cy} are constant weights.

4 Experiments

4.1 Evaluation protocol and implementation

Every paper employs different metrics, datasets, and tasks in the aging literature. Therefore, we include a large set of metrics, datasets, and tasks in our experiments to allow comparison with most existing methods.

Datasets. In this paper, we employ three widely-used, publicly available high-resolution datasets for face aging and analysis:
• *FFHQ-RR*: Initially proposed in [43], this aging dataset based on FFHQ [15] comprises of 48K images depicting people from 20 to 69 years old. Because of this *Restricted age Range*, we refer to this dataset as *FFHQ-RR*. Images are downsampled to 224×224.
• *FFHQ-LS*: Tis aging dataset, introduced in [26], is composed of the 70K images from FFHQ [15], manually labeled in 10 age clusters that try to capture both geometric and appearance changes throughout a person's life: 0-2, 3-6, 7-9, 10-14, 15-19, 20-29, 30-39, 40-49, 50-69 and 70+ years old. Consequently, this dataset is referred to as *FFHQ-LS* because of its *LifeSpan* age range. The resolution of these images is 256×256 pixels.
• *CelebA-HQ* [13, 22]: It consists of 30K images at 1024×1024 resolution, which we downsample to 224×224 pixels. The only age-related label in the dataset is *young*, which can be either true or false.

The use of *FFHQ-RR* and *FFHQ-LS* may seem redundant since they are both based on the FFHQ dataset, but we perform distinct experiments on both datasets to allow comparison with existing state-of-the-art methods (which report results on at least one of them).

Tasks. We employ two tasks to evaluate the performance:
• *Young* → *Old*: as in [43], we sample 1000 images belonging to the "young" category and translate them to a target age of 60. This task is only performed on CelebA-HQ.
• *Age group comparison*: similarly to [23], we consider different age groups: (20-29), (30-39), (40-49), and (50-69) on *FFHQ-RR* and additionally (0-2), (3-6), (7,9), (15,19) on *FFHQ-LS*. We again sample the first 1000 test images and translate every one of them into the central age of each of the four different age groups (25, 35, 45, and 55, respectively).

Metrics. We choose metrics to evaluate the two main aspects of the aging task. Firstly, the translated/generated images must preserve the content of the input image in terms of identity, facial expression, and background. Secondly, the age translation might be

Fig. 3: Comparison with State-of-the-art on CelebA-HQ for the *Young* $\rightarrow$ *Old* task employing a target age of 60 years old.

accurate. In particular, we adopt the following metrics:
• *LPIPS* [44] measures the perceptual similarity when the target age coincides with the input image age.
• *Age Mean Absolute Error (MAE)*. We employ a pretrained and independent age estimation network to compare the predicted age with the target age given an input image. As we already use the DEX pretrained classifier [33] at training time, we utilize Face++ API [4]. Experiments show that DEX is more biased towards younger age predictions than Face++. Therefore, reporting the MAE to the input target age a_t would be biased. To compensate for this DEX-Face++ misalignment, we estimate the age of the original images with Face++ and compute the mean for each group. We then report the distance between the mean group predicted age and the transformed image predicted age.
• *Kernel-Inception Distance [4] (KID)* assesses that the generated images are similar to real ones for similar ages. While FID [9] is adopted in [23], we adopted KID as it is better suited for smaller datasets. We report the KID between original and generated images within the same age groups.
• *Gender*, *Smile*, and *Face expression* preservation and *Blurriness*: Face++ provides these metrics to evaluate input image preservation and quality. *Gender*, *Smile*, and *Face expression* preservation are reported in percentages as in [43].

Implementation details. We use the same training settings as StyleGAN2-ADA [14] with $\lambda_r = 10$, $\lambda_C = 0.06$, $\lambda_D = 1$, $\lambda_{cy} = 10$. The optimizer used is Adam with $lr = 0.0025$ and $\beta_1 = 0$, $\beta_2 = 0.99$. FFHQ-RR and CelebA-HQ models are trained for 65 epochs with a batch size of 18. FFHQ-LS is trained for 140 epochs with a batch size of 16. All experiments are run on a single Nvidia A100 GPU.

4.2 Comparison with State-of-the-Art

From our literature review (Sec. 2), we identify HRFAE [43] and LATS [26] as the two main competing methods. Indeed, Re-aging GAN [23] cannot be included in the

[4] Face++ Face detection API: `https://www.faceplusplus.com/` (last visited on July 20, 2022).

Fig. 4: Qualitative comparison with HRFAE. The images corresponding to the input ages are highlighted with red frames.

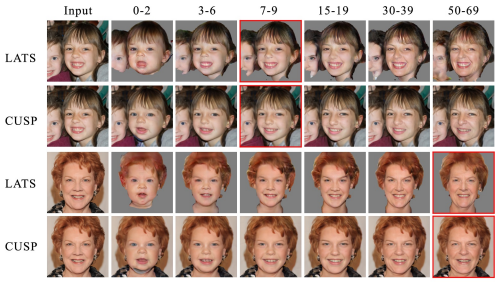

Fig. 5: LATS comparison for different age targets. The images corresponding to the input ages are highlighted with red frames.

Table 1: User study on four different aspects of image aging comparing CUSP.

	Age accuracy			*Identity preservation*			*Overall quality*			*Natural progression*
	20-29	*50-69*	*Added*	*20-29*	*50-69*	*Added*	*20-29*	*50-69*	*Added*	-
CUSP	**60.2**	**72.9**	**66.6**	**50.8**	**63.7**	**57.3**	**55.8**	**67.7**	**61.8**	**60.6**
HRFAE [43]	17.5	15.6	16.6	24.4	24.0	24.2	21.7	20.6	21.1	24.9
LATS [26]	22.3	11.5	16.9	24.8	12.3	18.5	22.5	11.7	17.1	14.5

comparison since neither the code nor the age classifier used for evaluation is publicly available. Since HRFAE and LATS report experiments on different datasets and follow different protocols, we perform experiments using the two tasks previously described. First, we follow HRFAE [43], which employs the *Young* $\rightarrow$ *Old* task on *CelebA-HQ*. In this case, the performance of FaderNet [19], PAG-GAN [41], IPC-GAN [40], and HRFAE (on 1024×1024 resolution images) is reported in [43] and is included in our experimental comparison. Second, we employ the *age group comparison* task to allow better comparison with LATS [26] on the most challenging *FFHQ-LS* dataset. Indeed, since no automatic quantitative evaluation is reported on the *FFHQ-LS* in [26], we chose the *age group comparison* task that provides richer analysis than the *Young* $\rightarrow$ *Old* task.

User study. We conducted a study on 80 different users comparing CUPSP with HRFAE [43] and LATS [26]) on the young-to-old and old-to-young tasks on FFHQ-RR. Similarly to [26], we asked about user preferences regarding identity preservation, target age accuracy, realism, the naturalness of the age transition, and overall preference.

As seen in Table 1, CUSP outperforms HRFAE [43] and LATS [26] in every single category by a large margin (CUSP was selected globally in 62% of cases, compared to 22% and 17%, respectively). Furthermore, CUSP's results depict people of the target age with greater accuracy while maintaining the source image identity. On top of that, it outputs higher quality images, and the progression seems more natural and realistic.

Qualitative comparison. In Fig. 3, we show a qualitative comparison with the state-of-the-art evaluated on the *celebA-HQ* dataset, where we transform the input image to the age of 60 years old. First, we observe that Fader, PAGGAN, and IPCGAN generate im-

Table 2: Quantitative comparison on *CelebA-HQ* for the *Young* $\rightarrow$ *Old* task employing a target age of 60. CUSP HP (High preservation) is run with $\sigma_m = \sigma_g = 1.8$.

Method	**Predicted Age**	**Blur**	**Gender**	**Smiling**	**Neutral**	**Happy**
Real images	*68.23 ± 6.54*	2.40	-	-	-	-
FaderNet	44.34 ± 11.40	9.15	97.60	95.20	90.60	92.40
PAGGAN	49.07 ± 11.22	3.68	95.10	93.10	90.20	91.70
IPCGAN	49.72 ± 10.95	9.73	96.70	93.60	89.50	91.10
HRFAE	54.77 ± 8.40	**2.15**	97.10	**96.30**	**91.30**	**92.70**
HRFAE-224	51.87 ± 9.59	5.49	**97.30**	95.50	88.30	92.50
LATS	55.33 ± 9.33	4.77	96.55	92.70	83.77	88.64
CUSP HP	**67.76 ± 5.38**	2.53	93.20	88.70	79.80	84.60

Table 3: Quantitative comparison with LATS on the FFHQ-LS dataset for the *age group comparison* task. CUSP CP (Custom preservation) and LP (Low preservation) are run with $(\sigma_m, \sigma_g) = (8, 4.5)$ and $(\sigma_m, \sigma_g) = (8, 8)$ respectively.

	Age MAE							**Gender Preservation (%)**						
	0-2	**3-6**	**7-9**	**15-19**	**30-39**	**50-69**	**Mean**	**0-2**	**3-6**	**7-9**	**15-19**	**30-39**	**50-69**	**Mean**
LATS	7.68	8.91	6.59	5.19	**8.23**	**5.73**	7.05	72.2	70.6	74.2	**93.7**	**93.9**	**93.9**	**83.1**
CUSP CP	6.89	**8.26**	7.67	6.70	10.67	10.86	8.51	**74.5**	69.3	**78.1**	88.3	92.1	85.9	81.4
CUSP LP	**6.49**	9.29	**5.59**	**4.99**	8.36	5.74	**6.74**	69.0	**76.0**	**78.1**	87.4	86.1	80.1	79.4

ages with important artifacts. On the contrary, HRFAE, LATS, and our approach generate consistent images with only minor artifacts. However, only CUSP produces images that correspond to the correct target age. Other methods generate images where people look younger than expected since they are unable to make suitable structural changes. Furthermore, LATS operates only in the foreground, requiring a previous masking procedure; for this reason, in Fig. 3, the outputs related to LATS display a constant gray background. In addition, CUSP can preserve identity and non-age-related details.

We also perform a qualitative comparison with the two main competitors: HRFAE on *FFHQ-RR* in Fig. 4 and with LATS on *FFHQ-LS* in Fig. 5. We show that CUSP achieves more profound facial structure modifications (*e.g.*, thin face shapes that grow wider and wrinkled skin) and hair color transformation. The age progression is smooth. Close ages produce almost identical pictures, but global age progression seems realistic and natural. Regarding LATS (Fig. 5), we see that we obtain similar performance while our method has four major advantages: (1) it operates directly on the entire image and deals with backgrounds and clothing; (2) it does not require an externally trained image segmentation network; (3) CUSP employs a single network while LATS uses a separate network for each gender; and (4) it offers user control as shown in our ablation study (see Sec. 4.3).

Quantitative comparison. In Table 2, we report a quantitative comparison evaluated on the *CelebA-HQ* dataset employing the *Young* $\rightarrow$ *Old* task. Every model has been trained

on *FFHQ-RR*. Regarding HRFAE, we report the performance obtained with models trained and tested at 224×224 and 1024×1024 resolutions (referred to as HRFAE-224 and HRFAE, respectively). We used the available code for LATS to train a model on this dataset. We also report (first row) the mean age predicted by the Face++ classifier when feeding the images of the age class 60 according to the DEX classifier used at training time. We observe an 8.23-year discrepancy. In other words, to generate images that look similar to those labeled as 60 at training time, we need to predict images that the Face++ classifier will perceive on average as 68.23 years old. These experiments confirm that CUSP outperforms other methods, being the only method that substantially modifies the image to adjust the person's target age.

In addition, CUSP ranks second in terms of Blur, quantifying the good quality of our images. For instance, the performance of HRFAE-224 worsens the predicted age with respect to its 1024×1024 counterpart and deteriorates noticeably in the Blur metric, suggesting a severe drop in the generated image quality. Interestingly, the more profound and realistic transformations yielded by CUSP and LATS imply slightly worse scores according to the preservation metrics. Indeed, preservation metrics suffer from the increased ability to make structural changes to pictures. However, this drop in quantitative fidelity is not manifested in the user study or qualitative results (Figs. 5 and 4). Two hypotheses can explain this discrepancy between qualitative and quantitative results. First, several biases can impact the results (*e.g.*, sports clothing is replaced for formal clothes at higher ages, and glasses appear in older targets as well). In addition, there may also be some expression-related biases in different age groups. Second, the CUSP module more frequently targets the image's mouth and eye areas. Those areas are the most related to facial expression detection, and their blurring might negatively affect facial expression preservation.

We report in Table 3 a comparison with LATS, both trained and evaluated on the *FFHQ-LS* dataset. The results support the qualitative analysis performed regarding Figs. 4 and 5. Our proposed method is on par with LATS performance concerning the aging task and achieves those results while preserving numerous image details. CUSP with low preservation even outperforms LATS in terms of Mean Age-MAE. We also notice that our approach obtains similar performance in terms of gender preservation while employing a single model and not using gender annotations as in [26].

4.3 Ablation study

Architecture ablation. We consider four variants of our approach where we ablate the skip connections and the style encoder[5]. In (i), the style encoder is not used; an *Average Pooling layer*replaces E_s on top of the output from E_c. *(ii)* employs a style encoder but no skip connections, while *(iii)* employs skip connections in every layer. Finally, *(iv)* follows the proposed architecture employing skip connections in the second-to-last layer only. In order to make an unbiased evaluation of the architecture and not the masking operation performed by CUSP, we report the performance of CUSP with high preservation $(\sigma_m, \sigma_g) = (7.1, 0.0)$, as *(ii)* applies no masking.

Table 4: Ablation study: impact of the skip connections (*SC.*) and the style encoder.

	LPIPS	Age MAE	Mean KID
(i) No style encoder	0.84	6.21	0.0163
(ii) No *SC.*	1.70	**6.17**	0.0109
(iii) SC. at every layer	1.85	6.34	0.0175
(iv) Full	**0.78**	6.29	**0.0089**

Table 5: Ablation study: impact of the masking strategy used in CUSP.

	LPIPS	Age MAE	Mean KID
Top-class GB	1.25	6.19	0.0145
Class-indep. (Ours)	**0.78**	**6.29**	**0.0089**

Results shown in Table 4 suggest that a separate style encoder, as in our *Full* model *(iv)*, yields better reconstruction (lower LPIPS) and similar aging performance (Age MAE and Mean KID) than using a single encoder for both content and style as in *(i)*. Regarding skip connections, not using them leads to an important reconstruction error (see high LPIPS) since the network cannot reconstruct the image details. However, skip connections in every layer also results in low reconstruction performance. We hypothesize that the model faces optimization issues. More specifically, adding skip connections on every layer dramatically increases the decoder's complexity (approximately doubling its number of parameters), making the network slower and harder to train.

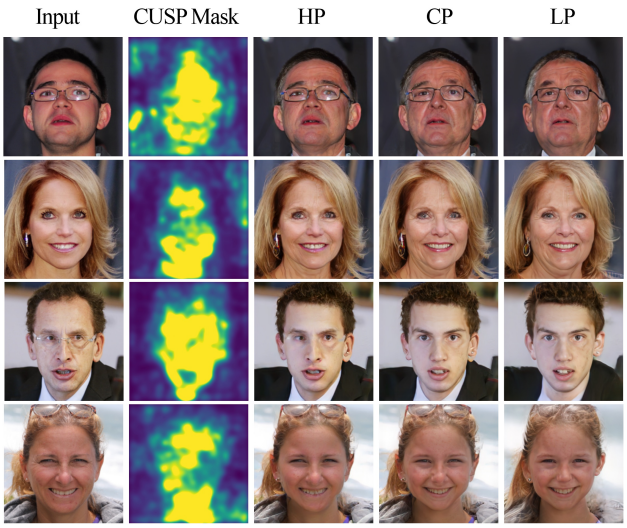

Fig. 6: Impact of the kernel value: images obtained with High, Low, and Custom structure preservation (LP, HP, and CP). HP:$(\sigma_m, \sigma_g) = (0, 0)$; CP:$(\sigma_m, \sigma_g) = (9, 0)$; HP:$(\sigma_m, \sigma_g) = (9, 9)$. The second column shows the mask estimated by CUSP.

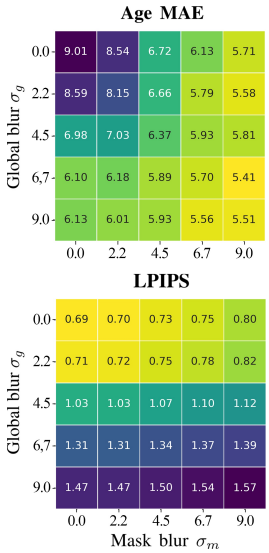

Fig. 7: CUSP parameters and impact on Age MAE (left) and LPIPS$\times 10$ (right).

CUSP module analysis. In Fig. 6, we qualitatively evaluate the impact of the kernel values used in CUSP. We compare images obtained with Low, Custom, and High structure preservation (referred to as LP, CP, and HP), where we use kernel values ranging from $\sigma = 0$ to $\sigma = 9$. We also display the mask $\mathbf{M}$ estimated by the CUSP module. We observe that when the user provides low kernel values (*i.e.*, higher preservation), the shape of the face is kept, while with higher kernel values, the network has the freedom to change its shape. The impact is clearly visible on the neck and chin of the women in the second and last row.

The visualization of the mask shows that our approach identifies those regions that change with age (chin, mouth, and forehead). We also quantitatively measure the impact of each kernel parameter. In Fig. 7, we report the Age MAE and LPIPS while changing the local and global blur parameters. By increasing the local blur, we can see that CUSP achieves a significantly lower age error while keeping a small reconstruction error. On the contrary, using global blur to improve the age performance (*i.e.*, reduce the age MAE) implies a substantial increase in the LPIPS metric, reflecting some loss of details. Overall, these experiments demonstrate the conflicting nature of aging and reconstruction performances. These observations further justify our motivation to offer the user the possibility of controlling this trade-off, thereby demonstrating the value of CUSP and its masking strategy. The ability to modify both σ_m and σ_g with different values allows us to achieve the same age-accurate transformation results while minimizing the reconstruction performance drop.

We complete this analysis with an ablation study regarding the GB-based computation of the CUSP masks. More precisely, two strategies are compared: in *Top-1 class*, we apply GB on the most-activated class, while in *class-independent*, we adopt the proposed strategy of taking the sum of the classification layer before softmax. Results reported in Tab. 5 demonstrate that the class-independent strategy performs best. Indeed, using every class output from the age classifier might benefit the masking, as every age-related feature is relevant for the translation, not only those involving its current age.

5 Conclusions

We present a novel architecture for face age editing that can produce structural facial modifications while preserving relevant details in the original image. Our proposal has two main contributions. First, we propose a style-based strategy to combine the style and content representations of the input image while conditioning the output on the target age. Second, we present a Custom Structure Preservation (CUSP) module that allows users to adjust the degree of structure preservation in the input image at inference time. We validate our approach by comparing six state-of-the-art solutions and employing three datasets. Our results suggest that our method generates more natural-looking, age-accurate transformed images and allows more profound facial changes while adequately preserving identity and modifying only age-related aspects. An extensive user study further confirmed this analysis. We plan to extend CUSP to other image editing tasks in future works.

References

1. Ak, K.E., Lim, J.H., Tham, J.Y., Kassim, A.A.: Attribute manipulation generative adversarial networks for fashion images. In: IEEE/CVF ICCV (2019)
2. 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)
3. Antipov, G., Baccouche, M., Dugelay, J.L.: Face aging with conditional generative adversarial networks. In: IEEE ICIP (2017)
4. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
5. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: IEEE/CVF CVPR (2018)
6. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Zhang, K., Tao, D.: Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In: IEEE/CVF CVPR (2019)
7. Fu, Y., Guo, G., Huang, T.S.: Age synthesis and estimation via faces: A survey. IEEE T-PAMI (2010)
8. He, Z., Kan, M., Shan, S., Chen, X.: S2gan: Share aging factors across ages and share aging trends among individuals. IEEE/CVF ICCV (2019)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Neurips **30** (2017)
10. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: IEEE/CVF ECCV (2018)
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE/CVF CVPR (2017)
12. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
13. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. ICLR (2017)
14. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. arXiv preprint arXiv:2006.06676 (2020)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE/CVF CVPR (2019)
16. 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)
17. Kemelmacher-Shlizerman, I., Suwajanakorn, S., Seitz, S.M.: Illumination-aware age progression. In: IEEE/CVF CVPR (2014)
18. Kim, D., Khan, M.A., Choo, J.: Not just compete, but collaborate: Local image-to-image translation via cooperative mask prediction. In: IEEE/CVF CVPR (2021)
19. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., DENOYER, L., et al.: Fader networks: Manipulating images by sliding attributes. In: Neurips (2017)
20. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: IEEE/CVF ECCV (2018)
21. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Neurips (2017)
22. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: IEEE/CVF ICCV (2015)
23. Makhmudkhujaev, F., Hong, S., Park, I.K.: Re-aging gan: Toward personalized face age transformation. In: IEEE/CVF ICCV (2021)

24. Miyato, T., Koyama, M.: cgans with projection discriminator. arXiv preprint arXiv:1802.05637 (2018)
25. Muhammad, M.B., Yeasin, M.: Eigen-cam: Class activation map using principal components. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2020)
26. Or-El, R., Sengupta, S., Fried, O., Shechtman, E., Kemelmacher-Shlizerman, I.: Lifespan age transformation synthesis. In: IEEE/CVF ECCV (2020)
27. Pan, H., Han, H., Shan, S., Chen, X.: Mean-variance loss for deep age estimation from a face. In: IEEE/CVF CVPR (2018)
28. Park, T., Zhu, J.Y., Wang, O., Lu, J., Shechtman, E., Efros, A., Zhang, R.: Swapping autoencoder for deep image manipulation. Advances in Neural Information Processing Systems **33**, 7198–7211 (2020)
29. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: Anatomically-aware facial animation from a single image. In: IEEE/CVF ECCV (2018)
30. 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)
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. Springer (2015)
32. Rothe, R., Timofte, R., Gool, L.V.: Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision **126**(2-4), 144–157 (2018)
33. Rothe, R., Timofte, R., Van Gool, L.: Dex: Deep expectation of apparent age from a single image. In: IEEE/CVF ICCV-W (2015)
34. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: IEEE/CVF ICCV (2017)
35. Siarohin, A., Sangineto, E., Lathuiliere, S., Sebe, N.: Deformable gans for pose-based human image generation. In: IEEE/CVF CVPR (2018)
36. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A.: Striving for simplicity: The all convolutional net. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings (2015)
37. Srinivas, S., Fleuret, F.: Full-gradient representation for neural network visualization. Advances in neural information processing systems **32** (2019)
38. Tang, H., Xu, D., Sebe, N., Yan, Y.: Attention-guided generative adversarial networks for unsupervised image-to-image translation. In: IJCNN (2019)
39. Wang, W., Cui, Z., Yan, Y., Feng, J., Yan, S., Shu, X., Sebe, N.: Recurrent face aging. In: IEEE/CVF CVPR (2016)
40. Wang, Z., Tang, X., Luo, W., Gao, S.: Face aging with identity-preserved conditional generative adversarial networks. In: IEEE/CVF CVPR (2018)
41. Yang, H., Huang, D., Wang, Y., Jain, A.K.: Learning face age progression: A pyramid architecture of gans. In: IEEE/CVF CVPR (2018)
42. Yao, X., Newson, A., Gousseau, Y., Hellier, P.: A latent transformer for disentangled face editing in images and videos. In: IEEE/CVF ICCV (2021)
43. Yao, X., Puy, G., Newson, A., Gousseau, Y., Hellier, P.: High resolution face age editing. In: IEEE ICPR (2021)
44. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
45. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: IEEE/CVF CVPR (2017)

46. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE/CVF ICCV (2017)
47. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Multi-modal image-to-image translation by enforcing bi-cycle consistency. In: Neurips (2017)