# A Latent Transformer for Disentangled Face Editing in Images and Videos: Supplementary Material

Xu Yao[1,2], Alasdair Newson[1], Yann Gousseau[1], Pierre Hellier[2]
[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2] InterDigital R&I, 975 avenue des Champs Blancs, Cesson-Sévigné, France
{xu.yao,anewson,yann.gousseau}@telecom-paris.fr, Pierre.Hellier@InterDigital.com

We provide further details on the experiments described in the main paper. Sec. 1 presents additional results of disentangled attribute manipulation and sequential attribute manipulation on real images of FFHQ dataset [3]. Sec. 2 provides additional results of facial attribute editing on videos collected from FILMPAC library [1] and RAVDESS dataset [5]. We show that our method handles disentangled and sequential facial attribute manipulation on images and videos. Please refer to our project page to view the facial attribute editing on videos: https://github.com/InterDigitalInc/latent-transformer/tree/master/image.

## 1. More results on real image editing

We provide additional results of disentangled attribute manipulation on real images in Figure 1, where only one attribute is modified at a time from the first projected image. Figure 2 presents additional results on sequential attribute manipulation. Here, we successively manipulate a list of attributes, meaning that each modification is performed on top of all previous modifications. We have trained a separate latent transformer for each of the 40 attributes in CelebA-HQ dataset [2]. Our method generates disentangled and identity-preserving manipulations for most of the attributes. We show some failure cases in Figure 5. When changing 'wavy hair', only slight changes appear in the hair. One possible reason is that the hair structures are controlled by the noise inputs in StyleGAN [4], while the pre-trained encoder [6] uses fixed noise inputs during training, which is a reasonable choice as the noise inputs have too many degrees of freedom to be reconstructed. In the case of 'wearing hat', we fail to generate a real hat. This attribute is very unbalanced in CelebA-HQ, so that it is difficult to learn the correct transformation.

## 2. More results on video manipulation

As mentioned in Sec. 5.2 of the main paper, we present additional facial attribute editing results on videos in Figure 3. Each sub-figure corresponds to a frame extracted from the corresponding video, in which the indicated attributes are modified. For each video, we edit two attributes sequentially, and generate disentangled manipulation results. For example, in Figure 3(c) when changing the person to woman, our method does not influence the attribute 'beard', despite the fact that it is correlated with gender. Besides, by varying the scaling factor progressively along the sequence, we achieve progressive attribute editing on videos. As shown by the video in Figure 4, we can simulate a progressive smiling process by smoothly varying the scaling factor. Overall, our method generates stable and consistent manipulation results on videos, provided that motion is not too strong. When there are quick changes of pose, we observe lighting or geometric artifacts. These artifacts are in fact due to the projection in the latent space, and therefore necessarily extend to the manipulated videos. As can be seen from the video in Figure 6, the manipulation during the first half of the video is realistic and consistent. But when the face turns to a side pose, the projected face is not well reconstructed and therefore neither is the manipulated face. This may be due to the limited reconstruction capacity of the pre-trained encoder and StyleGAN model when the pose is not frontal.

## References

[1] FILMPAC. Filmpac footage boutique library. https://filmpac.com, 2017. 1

[2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 1

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of*

*the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 1

[5] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 1

[6] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021. 1

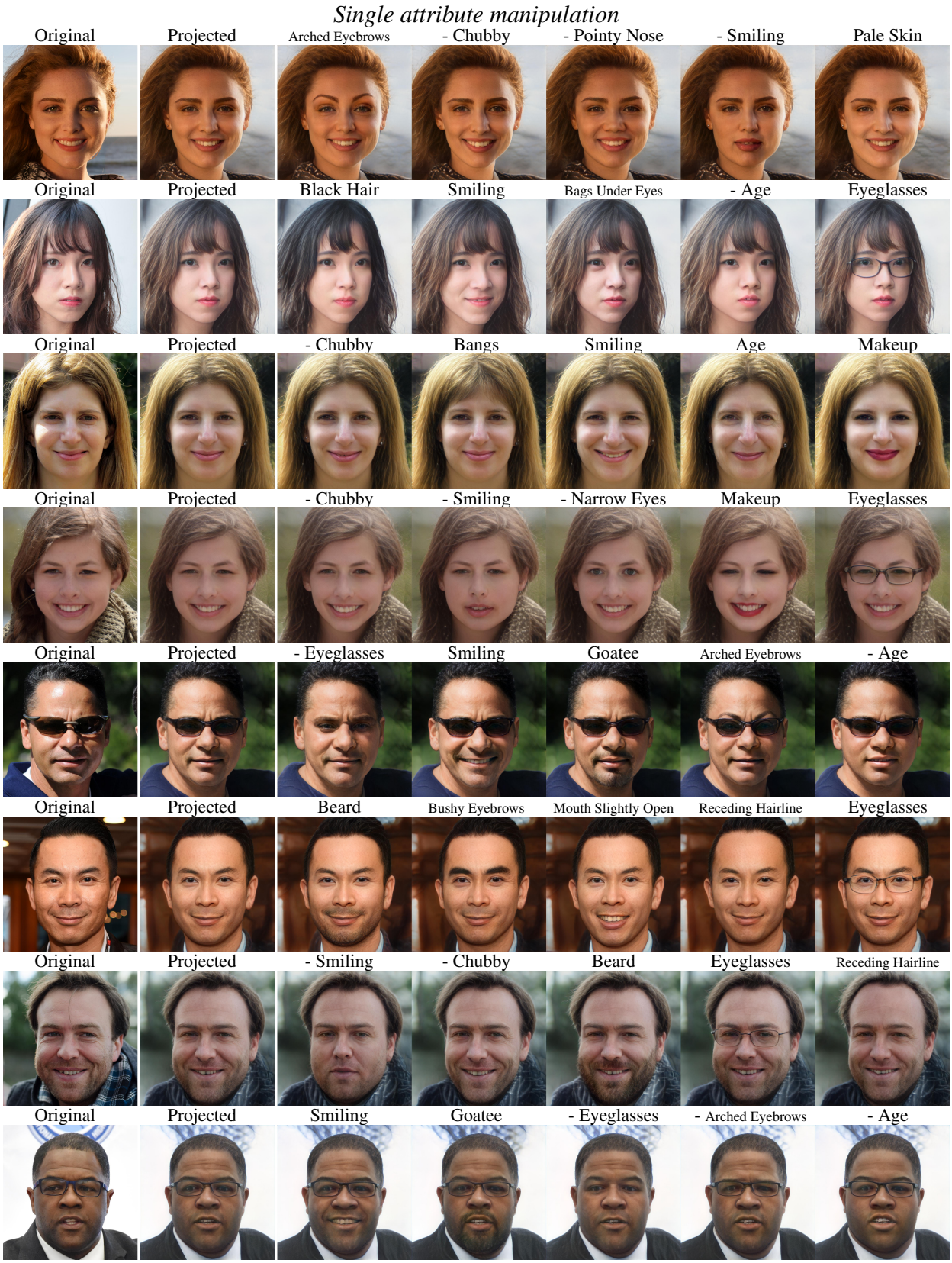



Figure 1: **Single attribute editing on real images**. Given an input image, we show manipulation results on several attributes, where each time a single attribute is modified from the first projected latent representation.


Figure 2: **Sequential attribute editing on real images**. Given an input image, we manipulate a list of attributes sequentially, where each time a single attribute is modified on top of the previous modfications.

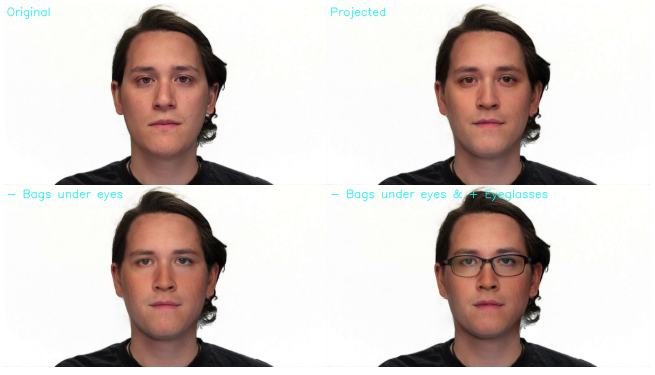
(a) 1_man_with_hat_Arched_Eyebrows_Beard.avi
- Arched Eyebrows, - Beard


(b) 2_woman_with_bricks_Eyeglasses_Age.avi
+ Eyeglasses, + Age


(c) 3_man_in_forest_Gender_Beard.avi
Gender, - Beard


(d) 4_man_with_muscle_Smiling_Young.avi
+ Smile, - Age


(e) 5_man_talking_Bags_Under_Eyes_Eyeglasses.avi
- Bags under eyes, + Eyeglasses




(f) 6_woman_turning_Smiling_Makeup.avi
- Smile, + Makeup


Figure 3: **Facial attribute editing on videos.** Each sub-figure corresponds to a frame extracted from the specified video, corresponding to the manipulation result of the indicated attributes. In each sub-figure, the upper row shows the original frame and the projected frame reconstructed with the encoded latent code in StyleGAN, the bottom row shows the manipulated frames for the first attribute and then for two attributes. Please open the video files to visualize the manipulation details.

7_woman_with_bricks_progressive_Smiling.avi, + Smile progressively


Figure 4: **Progressive attribute editing on videos.** By varying the scaling factor progressively along the sequence, the corresponding attribute is gradually varied. This figure show a frame extracted from the edited video, which corresponds to the progressive manipulation of the attribute 'smile'. From left to right: the original frame, the projected frame, and the manipulated frame. Please open the video file to fully visualize the manipulation.

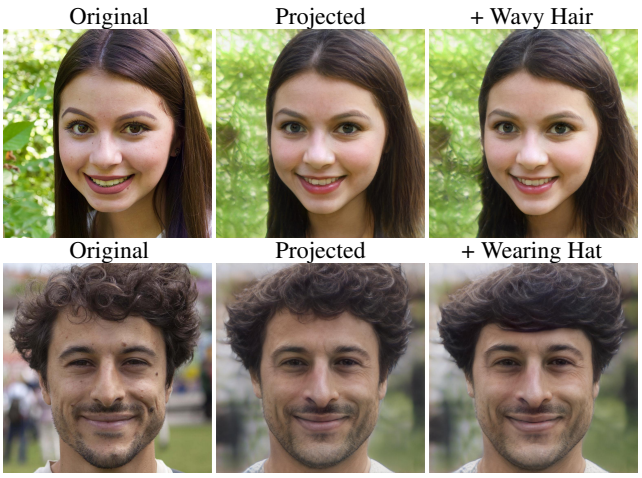



Figure 5: Failure case of attribute manipulation on real images. Each row corresponds to the manipulation of an attribute. From left to right: the original image, the projected image and the manipulated image. For 'wavy hair', our model yields only slight changes. In the case of 'wearing hat', we fail to generate a real hat.

failure_case_woman_sitting_Makeup.avi, + Makeup


Figure 6: Failure case of attribute manipulation on a video. This is a side pose frame extracted from the named video, which is the manipulation result of the attribute 'makeup'. From left to right: the original frame, the projected frame, and the manipulated frame. The face is not well reconstructed in the projected frame, and consequently the manipulated output contains defects. This is due to the limited generation capacity of the pre-trained encoder and the StyleGAN generator.