Third Time’s the Charm? Image and Video Editing with StyleGAN3

Yuval Alaluf, Or Patashnik et al.
Tel-Aviv University, Hebrew University, Adobe Research

Presenter : Taeu Kim

Contents

Problem

- The texture-sticking phenomenon in StyleGAN1 and StyleGAN2

  → hinders the temporal consistency and realism of generated and manipulated videos

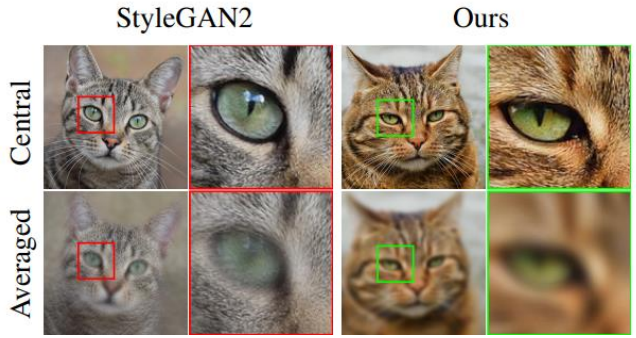
  → StyleGAN3

- Significant changes in StyleGAN3 raise many questions
  - 1. Disentanglement of its latent spaces

    → Compare StyleGAN3 with StyleGAN2
  - 2. Ability to accurately invert and edit real images

    → Do the techniques devised for identifying latent editing controls still work?

    → Which dataset-options are preferable for the editing : aligned or unaligned

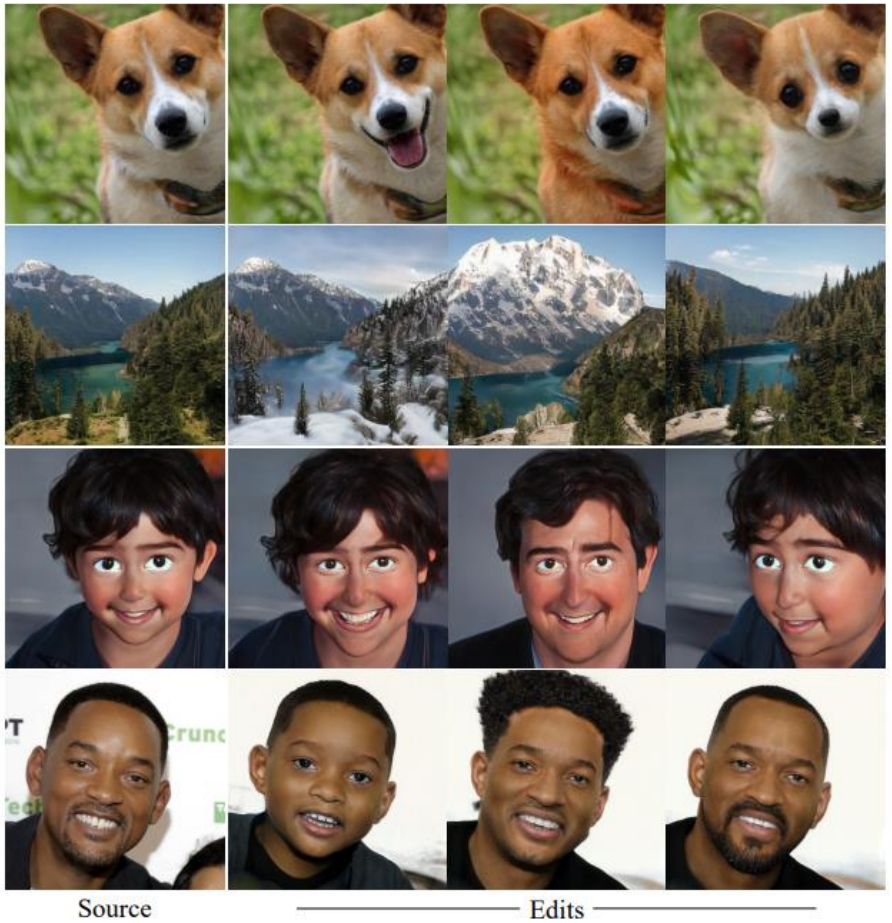


▲ Texture-sticking

Goal

Image & Video Editing with StyleGAN3

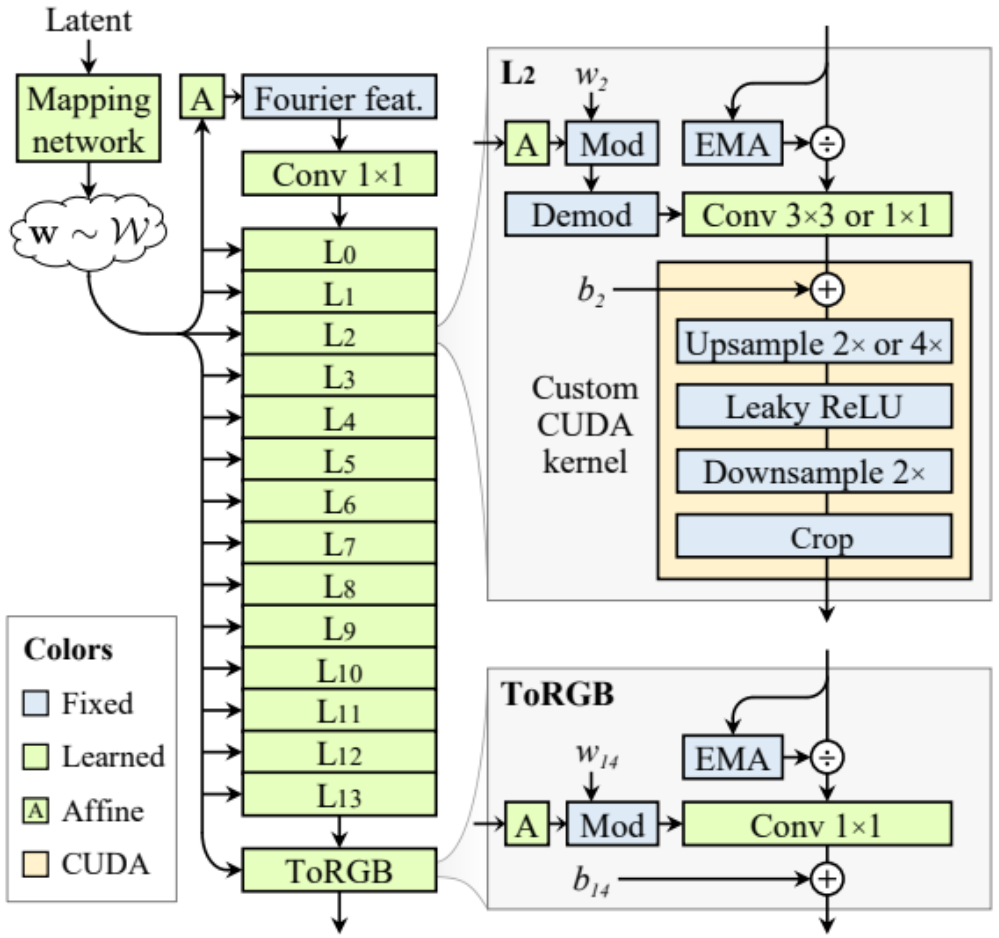
: leverage the capabilities of StyleGAN3 to reduce texture sticking and expand the field of view when working on a video with a cropped subject.



▲ Image Editing

▲ Video Editing

Analysis : 0. Architecture



▲ StyleGAN3 architecture

	StyleGAN2	StyleGAN3
Mapping Network (z to w)	-	Decreased depth
Constant input	4 x 4 constant	Fourier features that can be rotated and translated using four parameters (sin α, cos α, x, y)
Synthesis Network	Number of conv layers varies depending on an output resolution	A fixed number of conv layers (16)
Noise	Per-pixel noise	Eliminated
	Bilinear 2x upsampling, -	EMA, Custom CUDA kernel

Analysis : 1. Rotation Control

Figure 2. While StyleGAN3 trained on aligned data normally generates aligned images (leftmost image), translation and in-plane rotation can be controlled by applying an explicit transformation (r, t_x, t_y) over the Fourier features.

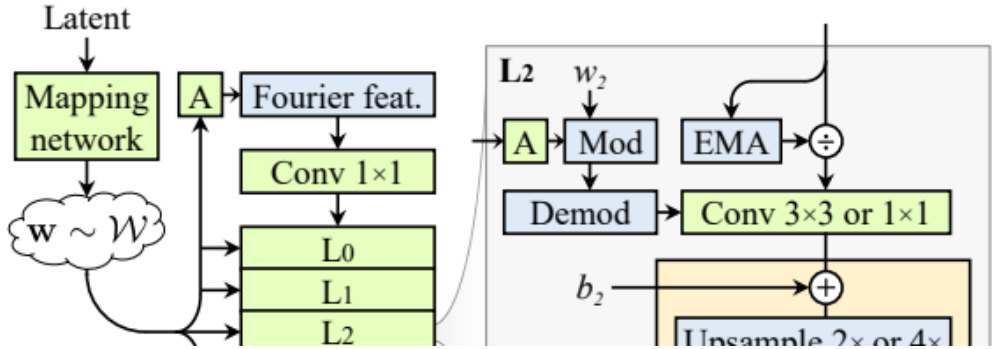


Figure 3. Pseudo-aligning images generated by unaligned generators. While these generators normally produce unaligned images (row 1), replacing w_0 with the average latent $\overline{w}$ yields roughly aligned images (row 2).

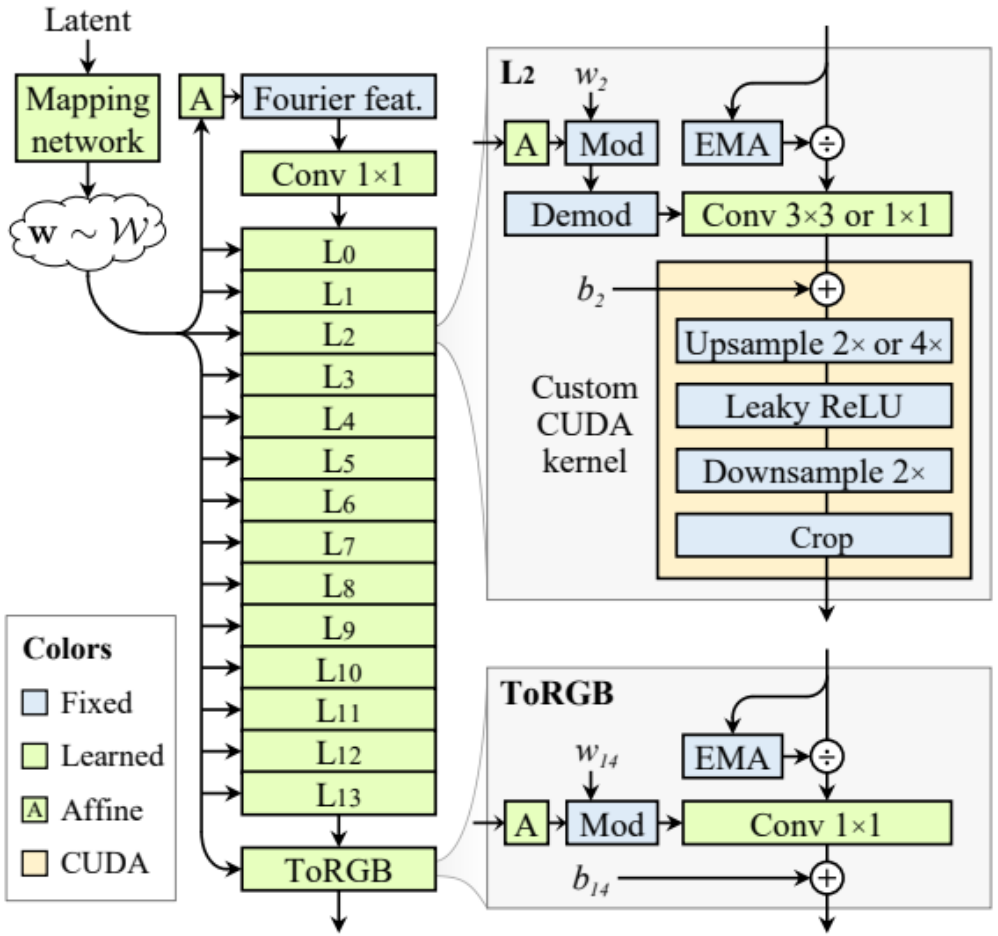
Analysis : 1. Rotation Control

Figure 4. The roles and entanglement of w_0 and w_1. Top row: altering only w_1 affects the in-plane rotation, as well as other visual aspects, implying they are entangled in w_1. Bottom row: holding w_0 and w_1 fixed while randomly sampling the remaining latent entries demonstrates that, for all practical purposes, the first two layers determine the translation and rotation.

Analysis : 2. Disentanglement

Generator	Space	Disent.	Compl.	Inform.
StyleGAN2	$\mathcal{Z}$	0.31	0.21	0.72
StyleGAN2	$\mathcal{W}$	0.54	0.57	0.97
StyleGAN2	$\mathcal{S}$	**0.75**	**0.87**	**0.99**
StyleGAN3 (A)	$\mathcal{Z}$	0.37	0.27	0.80
StyleGAN3 (A)	$\mathcal{W}$	0.47	0.43	0.94
StyleGAN3 (A)	$\mathcal{S}$	**0.89**	**0.76**	**0.99**
StyleGAN3 (UA)	$\mathcal{Z}$	0.36	0.26	0.80
StyleGAN3 (UA)	$\mathcal{W}$	0.45	0.41	0.94
StyleGAN3 (UA)	$\mathcal{S}$	**0.79**	**0.85**	**0.99**

Table 1. DCI metrics for StyleGAN2 and StyleGAN3. For both StyleGAN architectures, and for aligned and unaligned datasets, the DCI scores (disentanglement / completeness / informativeness) improve consistently from the initial Gaussian noise $\mathcal{Z}$, through the intermediate space $\mathcal{W}$, and to the style parameters $\mathcal{S}$, which control channel-wise statistics ($\mathcal{S} > \mathcal{W} > \mathcal{Z}$).



*StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation

Analysis : 3. Image Editing

Examine the effectiveness of various techniques for image editing with StyleGAN3, starting with the **W and W+ latent spaces**, and proceeding to S.

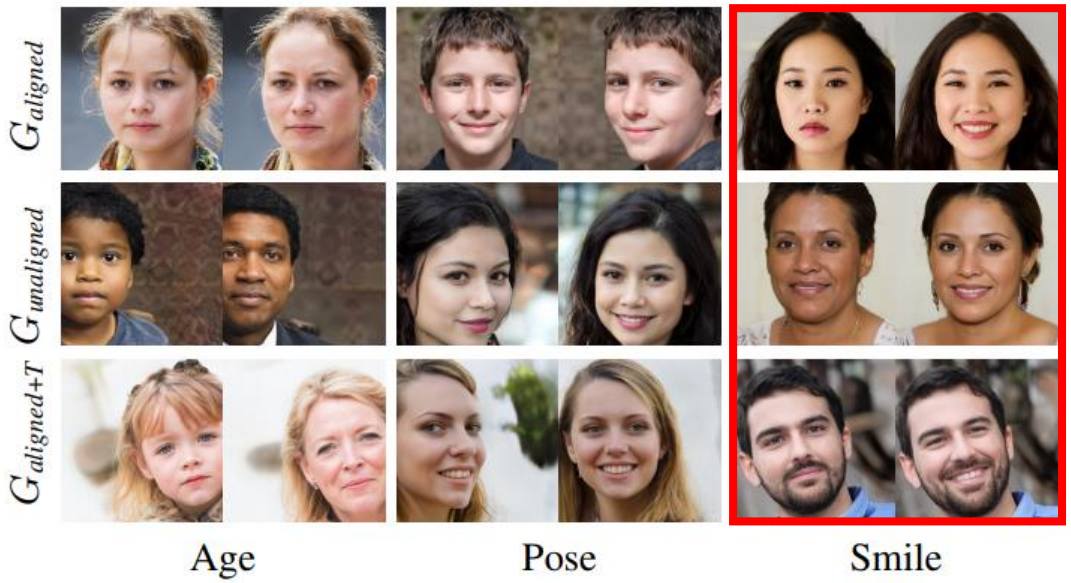


Figure 5. Linear editing in $\mathcal{W}$. Editing synthetic images using InterFaceGAN [60] directions in $\mathcal{W}$. Editing unaligned images can be done either using an unaligned generator $G_{unaligned}$, or using an aligned generator with an extra transformation $G_{aligned+T}$.

Figure 6. Non-linear editing in $\mathcal{W}+$. We edit images using the StyleCLIP mapping technique with StyleGAN3 trained on aligned faces. Even with non-linear editing paths, the edits are still entangled: local edits (e.g., expression/hairstyle) alter other attributes (e.g., background/identity).

Analysis : 3. Image Editing

Examine the effectiveness of various techniques for image editing with StyleGAN3, starting with the W and W+ latent spaces, and **proceeding to S.**

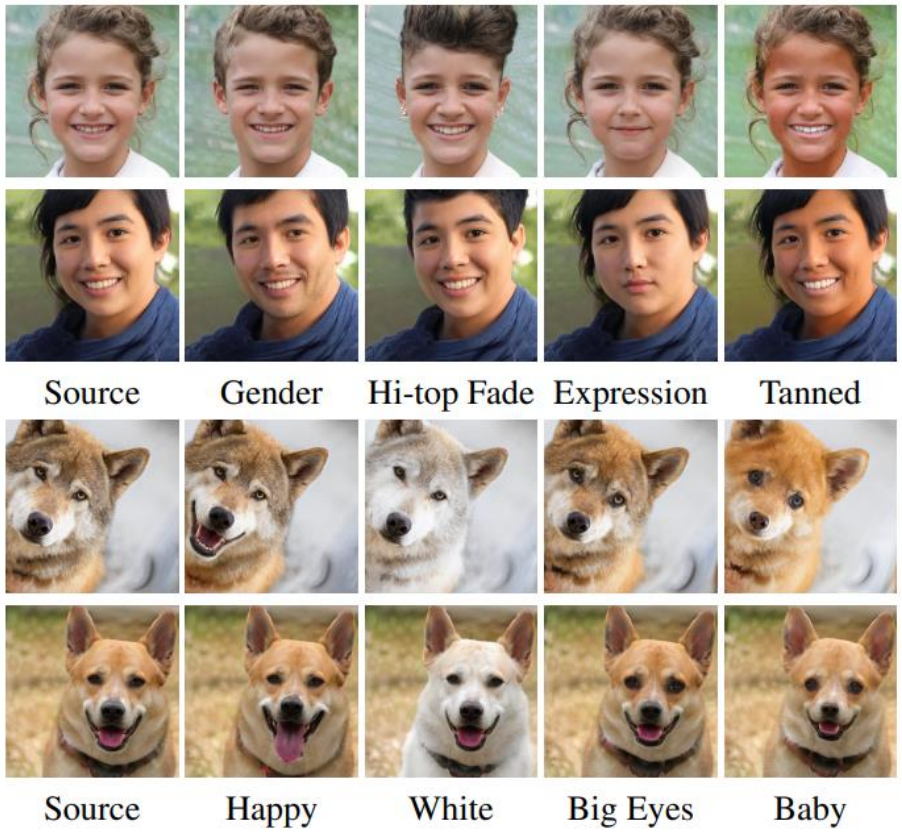
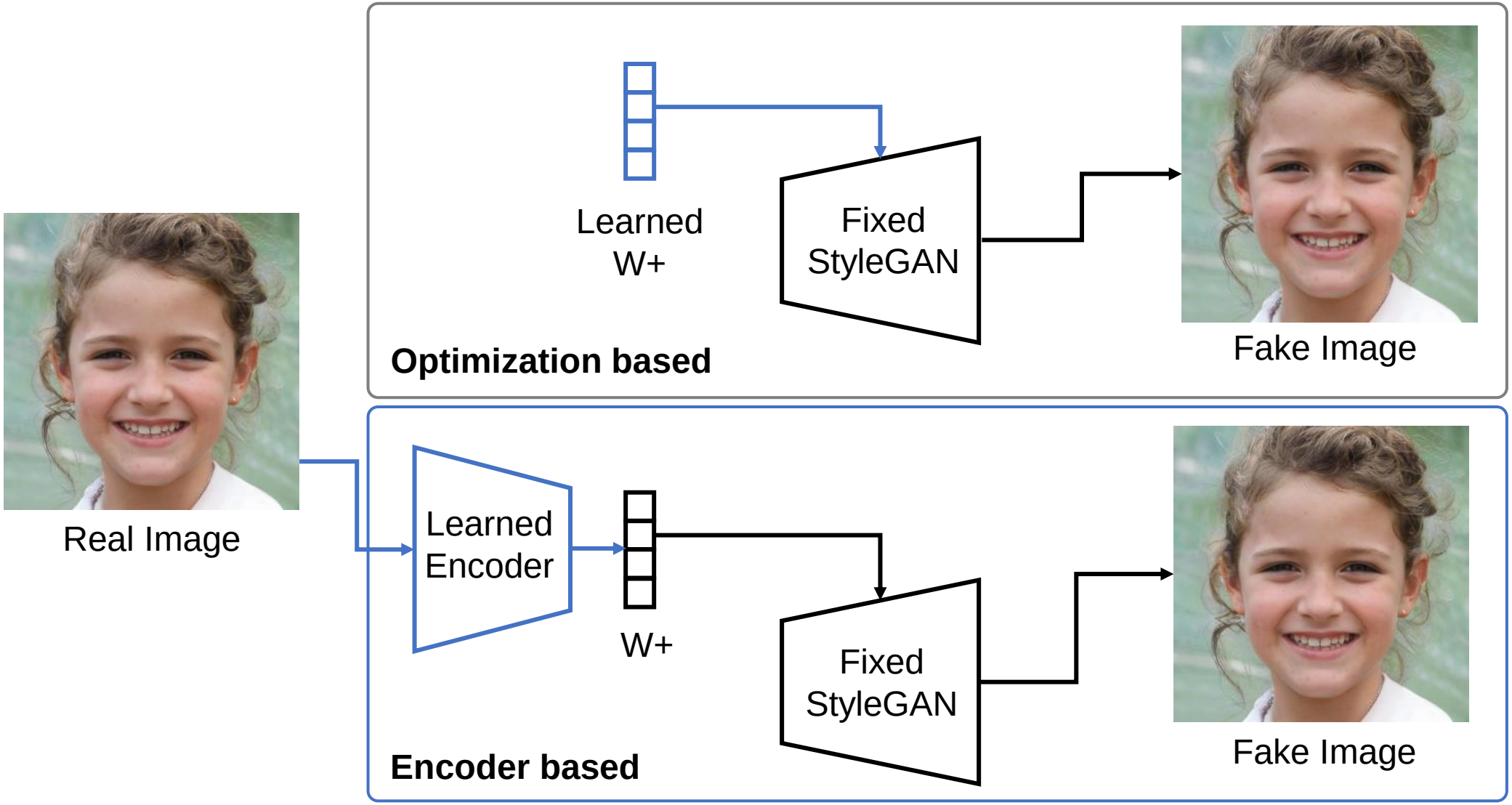
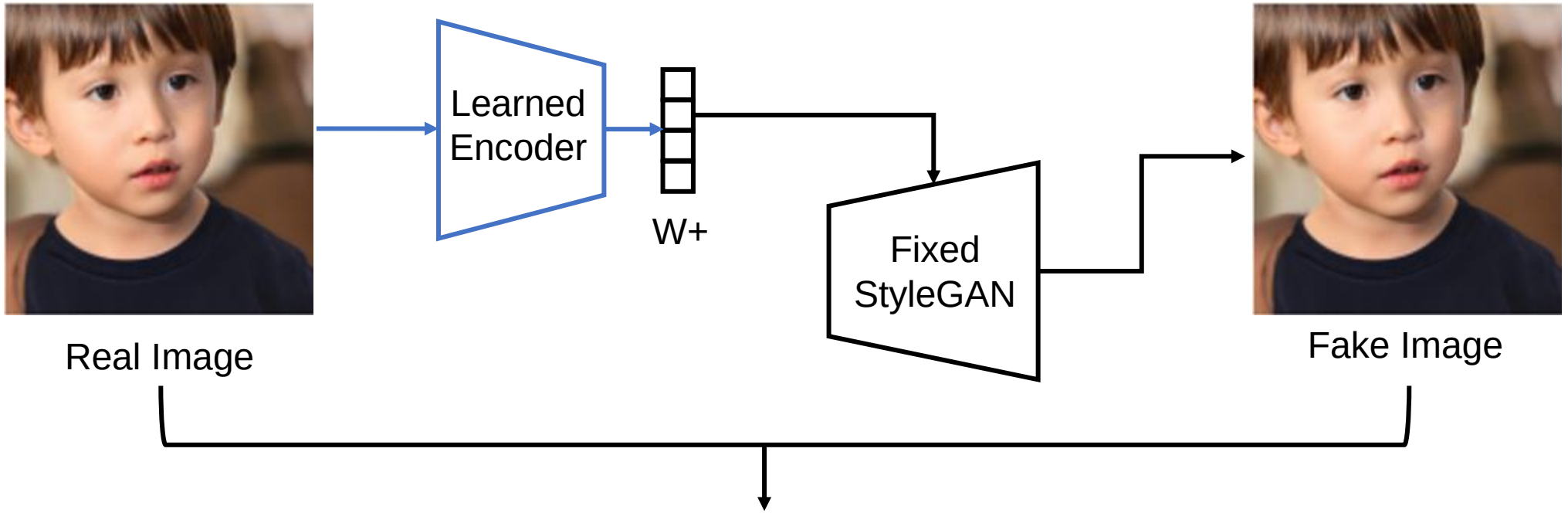


Figure 7. Editing in $\mathcal{S}$. We edit synthetic images using the StyleCLIP [48] global directions technique using StyleGAN3 generators trained on the FFHQ [35], AFHQv2 [13, 34], and Landscapes HQ [63] datasets.

Method : 1. Inversion



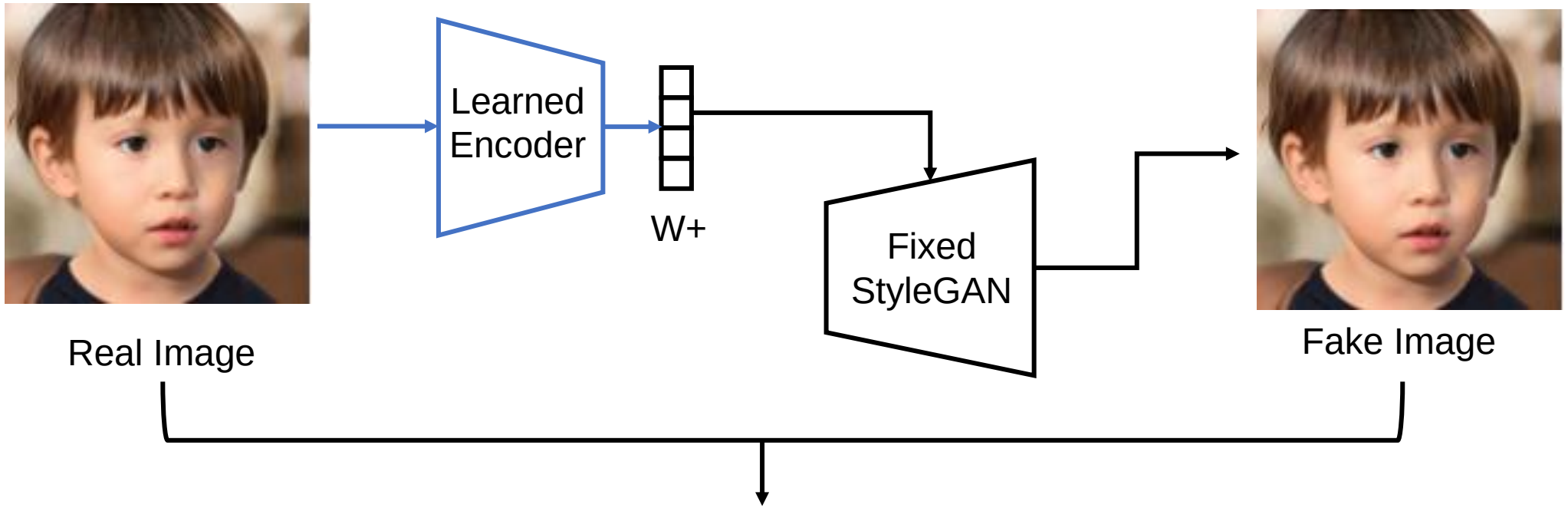
Method : 1. Inversion



Reconstruction Loss

1. L2
2. LPIPS
3. Identity Loss !

Method : 1. Inversion



Reconstruction Loss

1. L2
2. LPIPS
3. Identity Loss

Method : 1. Inversion

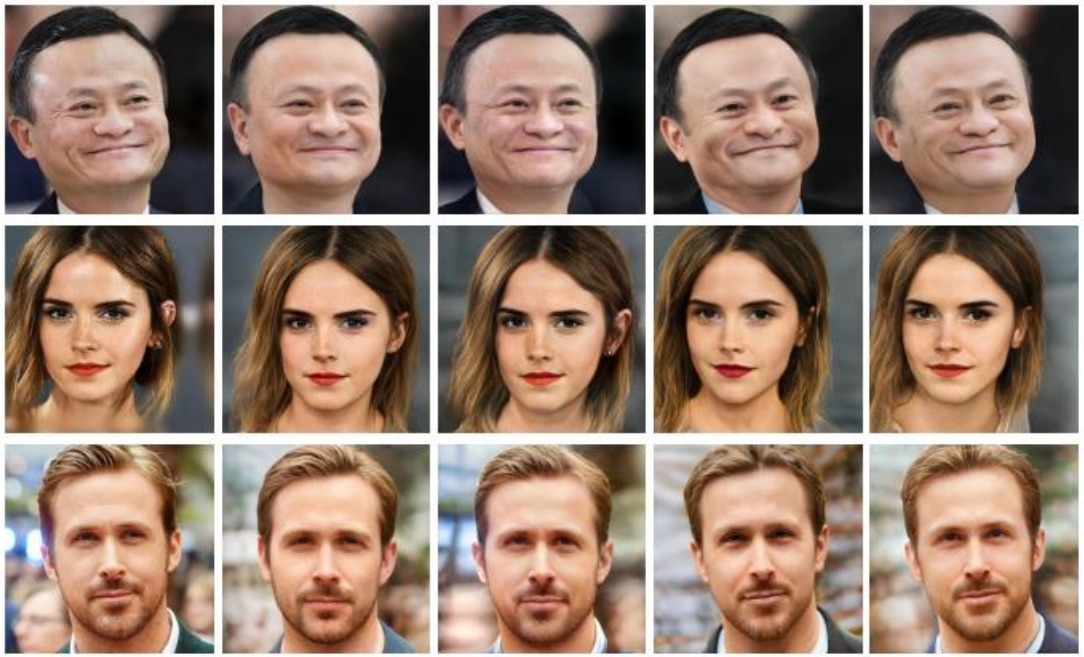


Figure 8. Reconstruction quality comparison between encoders trained for inverting StyleGAN2 and StyleGAN3 generators.

Method	↑ ID	↑ MS-SSIM	↓ LPIPS	↓ L_2	Time (s)
SG2 ReStyle$_{pSp}$	0.66	0.79	0.13	0.03	0.37
SG2 ReStyle$_{e4e}$	0.52	0.74	0.19	0.04	0.37
SG3 ReStyle$_{pSp}$	0.60	0.77	0.17	0.03	0.52
SG3 ReStyle$_{e4e}$	0.49	0.70	0.22	0.06	0.52

Table 2. Quantitative reconstruction results on the human facial domain measured over the CelebA-HQ [32, 44] test set. For StyleGAN3, we use a generator trained on aligned images.

Method : 1. Inversion and editing

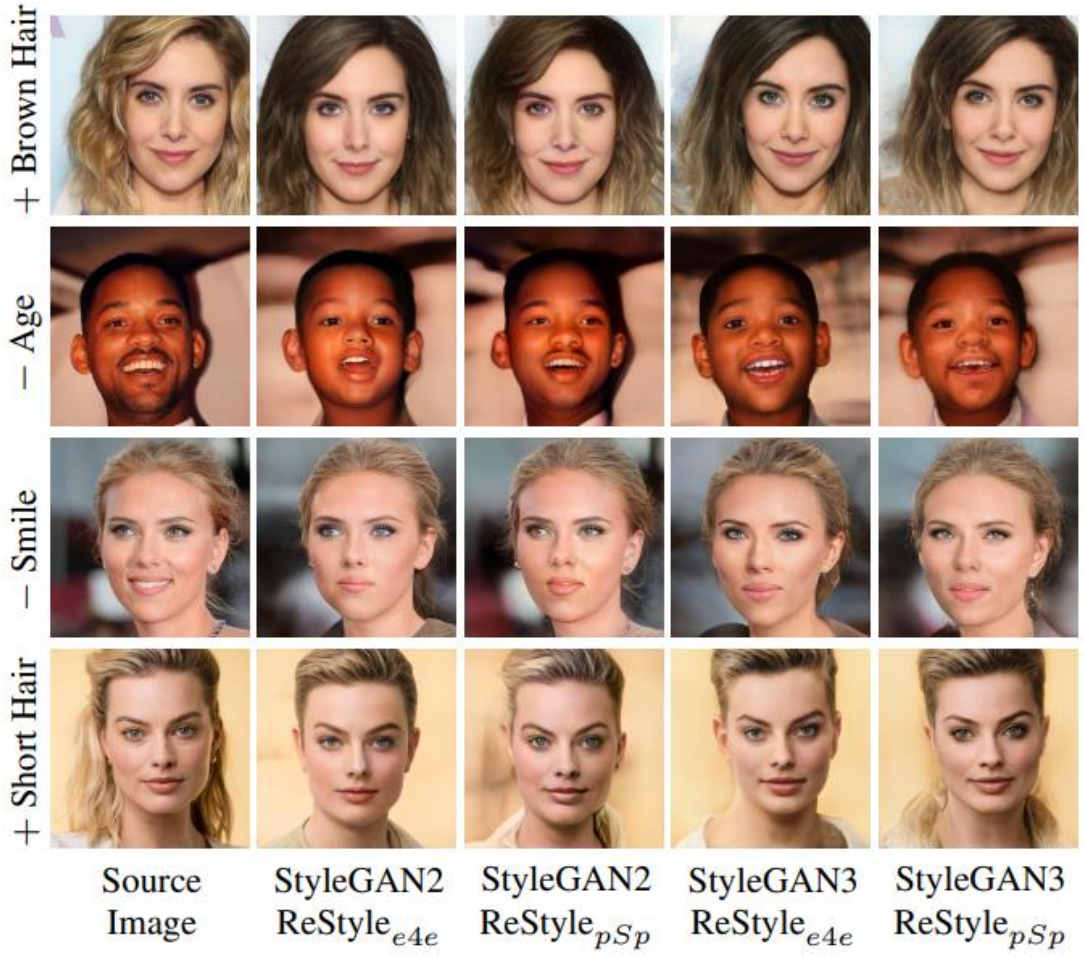
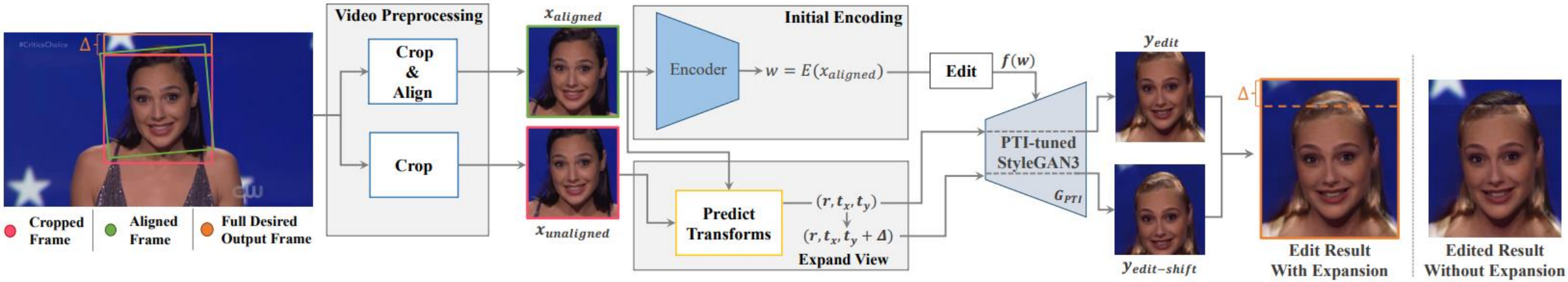
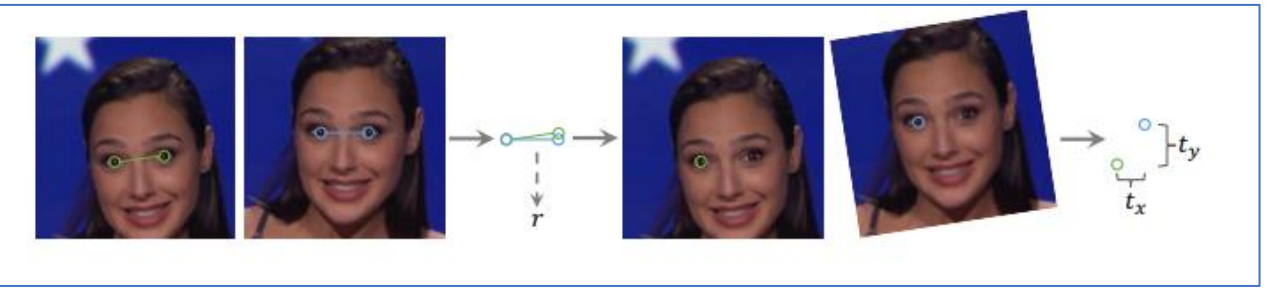


Figure 9. Editing comparison. We perform various edits [48, 60] over latent codes obtained by each inversion method.

Method : 2. Video





+ Predict Transforms

+ Latent Vector Smoothing

$$w_{i,smooth} = \sum_{j=i-2}^{i+2} \mu_j f(w_j)$$

$$T_{i,smooth} = \sum_{j=i-2}^{i+2} \mu_j T_j$$

+ Pivotal Tuning for Improved Reconstructions

$$y_{i,\mathrm{PTI}} = G_{\mathrm{PTI}}(w_i; (r_i, t_{x,i}, t_{y,i}))$$

Method : 2. Video

Figure 11. Sample results of our full video encoding and editing pipeline with StyleGAN3. Additional full video results and editing examples are provided in the supplementary materials.

Figure 12. Video reconstruction results obtained using our field of view expansion technique. In row 2 we provide the original unaligned reconstructions while in row 3 we provide the expanded video reconstruction.

Conclusion

1. the ability of StyleGAN3 to control the translation and rotation of generated images opens new intriguing opportunities.

2. StyleGAN3 latent space is somewhat more entangled than that of its predecessors.

   → We have shown that this may be alleviated by applying the inversion on aligned images and exploiting the transformation control to compensate for the alignments.

3. We have naturally focused on facial images and videos.

   → More research is required for other domains.