Supplementary Materials for Temporally consistent semantic video editing

Yiran Xu[1], Badour AlBahar[2], and Jia-Bin Huang[1]

[1] University of Maryland, College Park
[2] Virginia Tech

This supplementary material presents additional visual results and implementation details to complement the main paper.

1 Overview

We include the following contents as our supplementary material:

- This document, including the implementation details of our proposed approach, and additional results.
- Video results. We show our video results on the project page.

2 Implementation details

2.1 Datasets

RAVDESS. To evaluate human faces, we sub-sample 20 videos from the RAVDESS dataset [4]. To use GAN inversion, we apply the face alignment [3] with smoothing to each frame as a pre-processing. We first apply GAN inversion and different editing techniques to those frames. For *out-of-domain editing*, we use Restyle encoder [1] (PSP-based) and StyleGAN-NADA [2]. The resolution of GAN inversion output is 1024×1024, and the dimension of the latent code is 18×512. The pre-trained generators with different editing styles are directly from StyleGAN-NADA. For *in-domain editing*, we use Pivot Tuning Inversion (PTI) [6] on e4e [8] for inversion, and the StyleCLIP [5] mapper for editing. We train the editing directions, "eyeglasses" and "beard" use other directions from pre-trained mappers provided by the authors.

Internet Videos. The videos in the RAVDESS dataset have simple white background and close-up faces. Such videos do not reflect the complexity of faces in real videos and the potential challenges for achieving temporally coherent semantic editing. In light of this, we further show additional video results collected from the Internet for in-domain editing. We collected these Internet videos from YouTube and clipped them into 2-3 seconds. All the Internet videos are 720P with an FPS of 30.

For phase 1, we set the learning rate for MLP α_I to 0.005 and the number of optimization epochs to 5. For phase 2, we set the learning rate α_{II} to 0.0001,

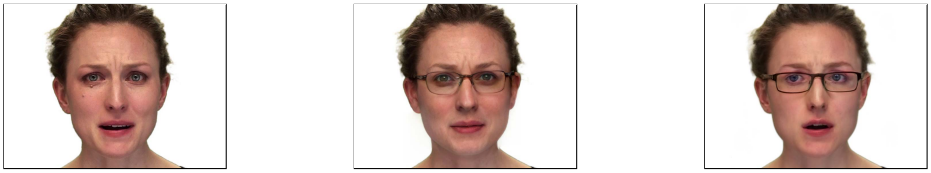

Fig. 1. Visual comparison with Latent Transformer (LT) [10]. LT cannot preserve the person's identity very well. Our method can preserve the identity and achieves a temporal consistent video.

and finetune G for 5 epochs. The regularization weight $\lambda_r = 100$. We use the same hyperparameters for all the collected Internet videos.

Optimization details. We use RAFT [7] as the flow network for computing dense 2D motion between the sampled, synthesized frames. We compute visibility masks using bi-directional consistency error maps, and we warp frames with flow field maps using bilinear interpolation. We use LPIPS loss [11] in the photometric loss. For the local regularization for the generator, we first sample a latent code z from a standard Gaussian distribution, and generate W_z using the StyleGAN's mapping network f: $W_z = f(z)$. To obtain a local code, we use linear interpolation between directly edited latent code W^{edit} and W_z following

$$W_r = W^{edit} + \alpha_{interp} \frac{W_z - W^{edit}}{||W_z - W^{edit}||_2}. \tag{1}$$

where α_{interp} is a parameter to control the step size of interpolation. The parameter α_{interp} controls the amount of attribute editing. We use $\alpha_{interp} = 30.0$ for all the experiments.

As described in the main paper, we propose a two-phase optimization. During the phase 1, we update only the latent code using an MLP and while keeping the generator parameters unchanged During the phase 2, we only optimize the generator and freeze all the latent code.

Note that the GAN inversion and the editing process can only be applied to *aligned and cropped faces*. We thus unalign the edited images back to the original video by implementing the stitch tuning method in [9] (since the source code was not available at the time).

2.2 MLP architecture

In phase 1, we train an MLP to predict the residual to achieve temporal consistency by updating the edited latent code. We use the same MLP architecture as StyleCLIP mapper [5]. The MLP has three groups (coarse, medium, and fine), following StyleCLIP mapper's design. We adjust the architecture based on its StyleCLIP editing direction for in-domain editing. For example, for "angry", "surprised", "eyeglasses", "beard", we remove the fine group; for "Johnny

Table 1. Comparison with (LT) Latent Transformer [10]

	$E_{warp} \downarrow$	
Editing categories	LT	Ours
eyeglasses	0.0066	0.0034
beard	0.0064	0.0032
Average perfomance	0.0065	0.0033

Depp", however, we use the full MLP architecture. For these in-domain editing, if there exists a pre-trained MLP mapper, we leverage the pre-trained mapper as the initialization. For out-of-domain editing, as we do not have pre-trained direction, we keep using an MLP containing only coarse and medium groups.

3 Additional results

3.1 Comparison with Latent Transformer [10]

On the RAVDESS dataset, We compare with Latent Transformer [10] with two semantic directions, "eyeglasses" and "younger", which are overlapped with ours. The results are reported in Table 3.1 and Figure 1

3.2 Additional visual results

For in-domain editing, we showcase additional results on Internet videos to present the method's capacity on real videos. For out-of-domain editing, we showcase results on selective RAVDESS [4] data. Please refer to our project page for more visual results.

(a) "Zombie" (b) "Rare pose"

Fig. 2. Limitations. From (a), it can be seen that earrings are added by GAN editing prior to our flow-based temporal consistency approach. Since our approach builds on existing GAN inversion and editing techniques, it will be affected by their quality. From (b), it can be seen that our method fails when there is a rare pose and a large motion.

3.3 Limitations

We show several limitations of our approach in Figure 2. First, our approach relies on plausible results from existing GAN inversion and editing techniques. We show an example of added earrings in Figure 2(a), and an example of a rare pose in Figure 2(a). Second, the GANs used in our experiments require the objects to be spatially aligned and thus may not yet be suitable for inverting and editing unconstrained videos. Third, our method relies on a high-quality GAN model that may be computationally expensive to train and often require diverse training images. Our full method (phases 1, 2 and 3) takes 40 minutes on a 150-frame video, on a single NVIDIA P6000 GPU.

References

1. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Restyle: A residual-based stylegan encoder via iterative refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2021) 1
2. Gal, R., Patashnik, O., Maron, H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators (2021) 1
3. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: ECCV (2020) 1
4. Livingstone, S.R., Russo, F.A.: 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, 3
5. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2085–2094 (October 2021) 1, 2
6. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. arXiv preprint arXiv:2106.05744 (2021) 1
7. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. arXiv preprint arXiv:2003.12039 (2020) 2
8. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG) **40**(4), 1–14 (2021) 1
9. Tzaban, R., Mokady, R., Gal, R., Bermano, A.H., Cohen-Or, D.: Stitch it in time: Gan-based facial editing of real videos. arXiv preprint arXiv:2201.08361 (2022) 2
10. Yao, X., Newson, A., Gousseau, Y., Hellier, P.: A latent transformer for disentangled face editing in images and videos. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13789–13798 (2021) 2, 3
11. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 2