

ConsistentAvatar: Learning to Diffuse Fully Consistent Talking Head Avatar with Temporal Guidance – Supplemental Document –

A IMPLEMENTATION DETAILS

Evaluate temporal consistency with optical flow. Optical flow is a computer vision technique used to quantify pixel motion between adjacent frames in a video sequence. Its fundamental principle involves estimating pixel motion by comparing the grayscale values of pixels between neighboring frames. Based on this, we utilize optical flow to measure the variation between generated frames of portrait videos, considering the typical variations between adjacent frames in videos. We use ground truth as a reference for comparison.

Given a test RGB video containing K frames $\{I_1, I_2, \dots, I_K\}$. First, convert the adjacent frames to grayscale images for subsequent calculations. Then, detect feature points in the current frame, such as corners or edges. These feature points should have similar positions in the adjacent frames. Finally, for each feature point, calculate its displacement in the adjacent frames, which is the optical flow vector. This process can be represented by the following formulas:

$$\{V\}_k = O(I_k, I_{k+1}), \quad (1)$$

where O represents the optical flow function, and $\{V\}$ represents the collection of obtained motion vectors. Next, we visualize the computed optical flow vectors as an image to intuitively observe the pixel changes between adjacent frames. We use color coding to represent the direction and magnitude of the optical flow, and we take the average pixel value of the resulting visualization image as an indicator of temporal consistency. As shown in Fig. 1, the visualization images obtained by our method are close to the ground truth, indicating good temporal consistency.

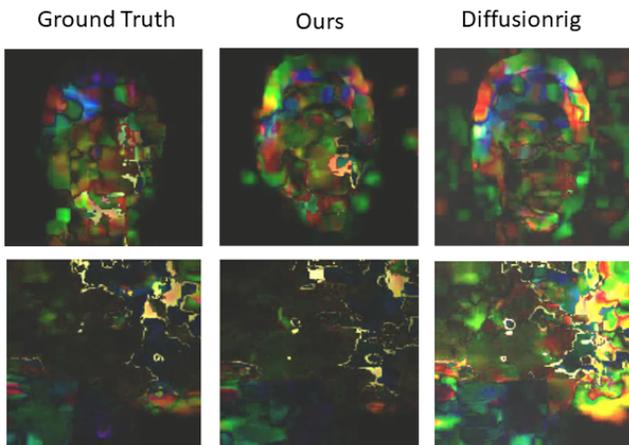


Figure 1: The visualization results of randomly selected two frames, evaluated for temporal consistency using optical flow.

B COMPARATIVE EXPERIMENTS

Diffusion models beat GANs. When discussing image generation using diffusion models, it is impossible to avoid mentioning Generative Adversarial Networks (GANs), both of which have demonstrated powerful capabilities in the field of image generation. As mentioned in article [1], GAN models can achieve high-quality generation, but this quality often comes at the expense of diversity. Moreover, the design of GAN models requires precise parameter selection, otherwise they are prone to instability, which limits their application in downstream tasks. Based on this, we believe that in the Temporally Consistent Module (Stage2), it is necessary to empirically demonstrate the advantages of diffusion models in generating Temporally-Sensitive Detail (TSD) maps compared to GANs. As shown in Fig. 2, we have designed a GAN-based architecture for generating TSD.

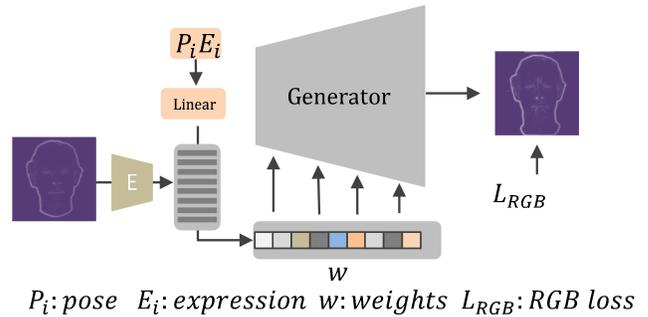


Figure 2: The GAN architecture used for generating TSD.

We compare the TSD learned through the diffusion model with the TSD learned through GAN for temporal consistency. We then use them as conditional inputs for ControlNet, comparing the temporal consistency of the generated videos (See Fig. 3).

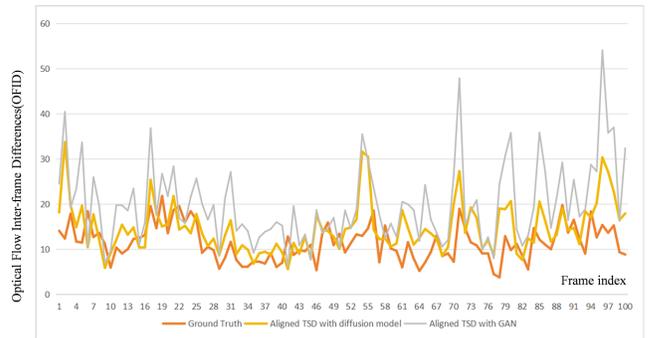


Figure 3: The results of aligning TSD using GAN and diffusion models respectively.

REFERENCES

- [1] Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. arXiv:2105.05233 [cs.LG]