CoDeF: Content Deformation Fields for Temporally Consistent Video Processing
Supplementary Materials

Anonymous CVPR submission

Paper ID 7925

Table 1. **Reconstruction PSNR** on DAVIS [2]

Video Name	LNA [1]	CoDeF
Blackswan	29.92	31.51
Boat	31.51	34.13
Car-turn	28.35	30.69
Kite-surf	28.37	34.26
Libby	28.29	30.05
Motorbike	29.85	32.99

The supplementary materials are structured as follows: We first outline the detailed configurations of the training process. Subsequently, we present additional quantitative results pertaining to the reconstruction. Lastly, we supply further qualitative results, expanding on various applications and a diverse array of video sequences.

1. Implementation Details

Our training networks comprise two MLPs equipped with HashEncoding. For the hash settings, we configure the number of features per level to 2 and set the level to 16. The base resolution starts from 16 and the log hashmap size is designated as 19. For the 3D deformable MLP, the scale per level is set to 1.38, and for the 2D MLP, it is set to 1.44. Pertaining to the MLP configurations, we employ eight layers for the 3D variant and two layers for the 2D variant. The activation function used in the MLP is ReLU. The initial learning rate is established at 1e-3 and is halved every 2,500 iterations. Our experiments' default parameters have the anneal begin and end steps set at 4,000 and 8,000, respectively. The total iteration step is limited at 10,000. The flow loss coefficient is set to 1, and the background loss is set at 0.03.

2. Quantitative Comparison

We follow the evaluation settings in **LNA** [1] and report more sequences in DAVIS dataset as shown in Table. 1

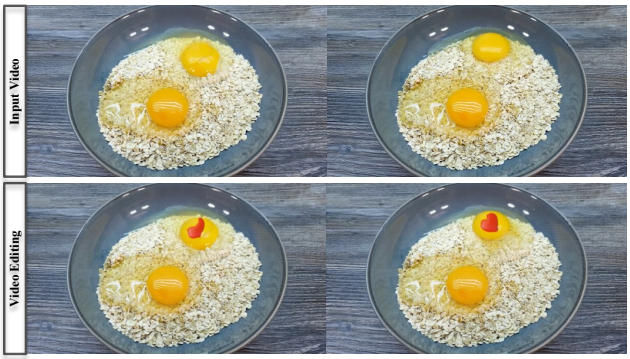

Figure 1. **User interactive video editing** achieved by editing *only one* image and propagating the outcomes along the time axis using our `CoDeF`. We strongly encourage the readers to see the supplementary videos to appreciate the temporal consistency.

3. More Results

In this section, we first show another application of CoDeF which is *User interactive Video Editing*. Our representation allows for user editing on objects with unique styles without influencing other parts of the image. As exemplified in Fig. 1, users can manually adjust content on the canonical image to perform precise edits in areas where the automatic editing algorithm may not be achieving optimal results.

We present further results generated using CoDeF for a variety of video sequences in Fig. 2, Fig. 3 and FIg. 4. Additionally, we supply videos that more effectively demonstrate the consistency within the videos. We strongly recommend that readers review these materials for a more comprehensive understanding.

References

[1] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Trans. Graph.*, 40(6):1–12, 2021. 1

[2] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbelaez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 1

Prompt: CG Style, Pink Hair

Prompt: Tifa from the Final Fantasy

Figure 2. **More results**

Prompt: Rainbow Smoke

Prompt: Chinese Ink Style

Figure 3. **More results**

Prompt: Earth

Prompt: Imaginary Cold Tune Butterfly

Figure 4. **More results**