StableVideo: Text-Driven Consistency-aware Diffusion Video Editing

Wenhao Chai[1], Xun Guo[2], Gaoang Wang[1], Yan Lu[2]

[1]Zhejiang University [2]Microsoft Research Asia

Introduction

Highlights

- Try to solve the consistency problem of diffusion video editing more efficient and effective.
- Propose a new video editing framework which can consistently manipulate the appearance of the objects.
- Achieve superior editing performance compared with SOTA approaches.

Motivation

Problem Statement

Given a video clip I, we obtain the mapping relationships of the atlases with respect to the pixel coordinate systems, named as $\mathrm{UV}^b(\cdot)$ and $\mathrm{UV}^f(\cdot)$, as well as the opacity α_i on the pixel coordinate, formulated as:

$$\mathrm{UV}_i^b(\cdot) = \mathcal{M}^b(I_i), \mathrm{UV}_i^f(\cdot) = \mathcal{M}^f(I_i), \alpha_i = \mathcal{M}^\alpha(I_i).$$

After that, we formulate the mapping from the atlas representation of the background A^b and the foreground A^f to the pixel coordinate systems of B_i and F_i:

$$B_i = \mathrm{UV}_i^b(A^b), F_i = \mathrm{UV}_i^f(A^f).$$

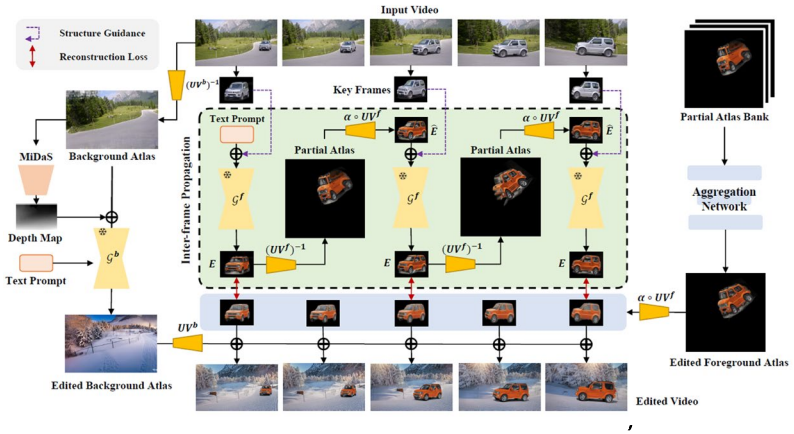
The entire video I and be reconstructed frame by frame:

$$I_i = \alpha_i \circ \mathrm{UV}_i^f(\mathcal{G}^f(A^f)) + (1-\alpha_i) \circ \mathrm{UV}_i^b(\mathcal{G}^b(A^b)),$$

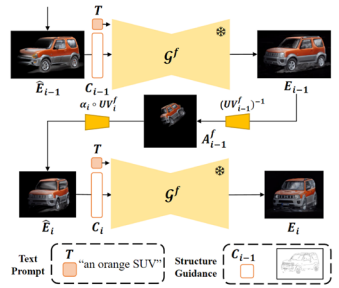
where $\circ$ denotes pixel-wise product.

Proposed Approach

StableVideo Framework



a) Inter-frame propagation

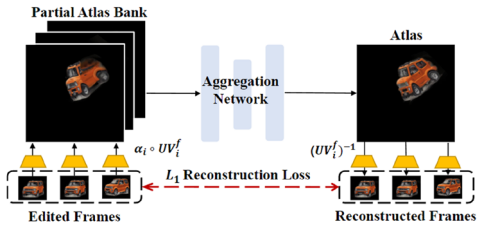


$$E_0 = \mathcal{G}^f(T, C_0)$$

$$A_{i-1}^f = (\mathrm{UV}_{i-1}^f)^{-1}(E_{i-1}),$$
$$\hat{E}_i = \alpha_i \circ \mathrm{UV}_i^f(A_{i-1}^f).$$

$$\hat{Z}_i(t_0) = \alpha(t_0)\hat{Z}_t(0) + \sigma(t_0)\mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$$

b) Aggregation Network



$$\mathcal{L}_{rec} = \sum_{i=1}^{N} ||E_i - \mathrm{UV}_i^f(A^f)||_1$$

Experiments

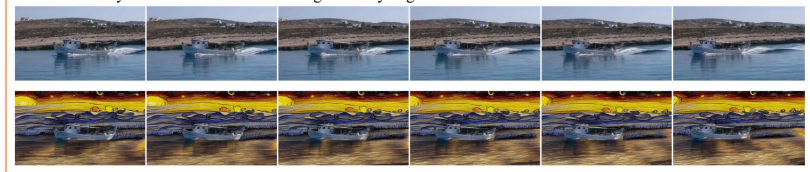
Compositing Editing

Foreground: a polar bear; Background: north pole.

Foreground: a car with graffiti; Background: Miami city.

Style Transfer

Both: in the style of Vincent Willem van Gogh's Starry Night.



Consistency & Complexity

Method	CLIP (↑)	LPIPS-P (↓)	LPIPS-T (↓)
Tune-A-Video	**0.2787**	0.6346	0.1851
StableVideo (*ours*)	0.2713	**0.1613**	**0.0386**

Method	Video Training	Edit Training	Edit Inference
Text2LIVE [1]	~ 10 hr	~ 1 hours	~ 10 sec
Tune-A-Video [4]	~ -	30 min	~ 4 min
StableVideo (*ours*)	~ 10 hr	-	~ 30 sec

Ablation Study