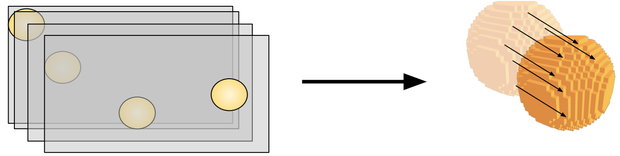
Temporally and Spatially Novel Video Frame Synthesis Using 4D Video Autoencoder

Authors: Bidipta Sarkar, Xinyi Wang, Feiyang (Kathy) Yu
Stanford University, Department of Computer Science

Introduction

We want to turn a video into a 4D scene



We generate **deep voxels**, a **camera trajectory**, and a **voxel flow**, which can be used for novel space-time view synthesis.

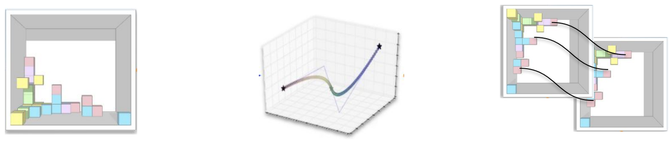
We build off of a pretrained static-scene Video Autoencoder and add flow encoder and flow decoder modules to capture **voxel flow.**

	D-NeRF	Video AE*	Deep Voxel Flow	Ours**
3D-Space	✅	✅	❌	✅
Dynamic	✅	❌	✅	✅
Feed-forward	❌	✅	✅	✅

Problem Statement

Input: video frames (T × C × H × W)

Encoder Output:
Voxels (C × H × W × D) **Cam Trajectory** (t × 6) **Flow** (H × W × D × 3)

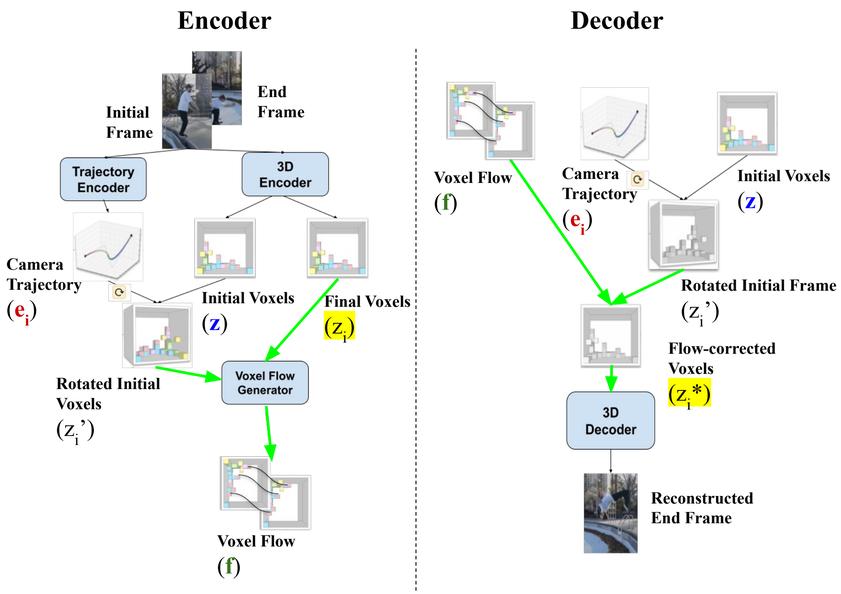


To evaluate, we compute the LPIPS, PSNR, and SSIM of novel space-time view synthesis outputs.

Dataset

HMDB51 "Stand" HMDB51 "Run"

Architecture



Loss:

$$\mathcal{L} = \cos\_\text{sim}(z^*_{i,\text{m}}, z_{i,\text{m}}) + \lambda_1 ||z^*_{i,\text{m}} - z_{i,\text{m}}||_1 + \lambda_{\text{perc}} \mathcal{L}_{\text{perc}}$$

- We apply this loss to interpolation and arbitrary-frame reconstruction
- To only compare relevant voxels, we create a saliency map from voxels to the final image to weight the importance of each voxel.

Qualitative Results

Middle Frame Synthesis
(From left to right: Ground Truth, Baseline (M0), M5)

Frame Reconstruction
(From left to right: Ground Truth, Baseline (M0), M5, Initial Frame)

Quantitative Results

Method	LPIPS (↓)	PSNR (↑)	SSIM (↑)	Ext. Test
M0	1.582	23.167	0.643	T
M1	1.556	23.392	0.652	T
M2	1.560	23.416	0.652	T
M3	1.560	23.455	0.653	T
M4	1.889	21.858	0.586	T
M5	**1.503**	**24.010**	**0.682**	T

Note: Models were tested on unseen HMDB51 action category.

Analysis

- Best Model [M5]
  - 5.0% improvement in LPIPS
  - 3.6% improvement in PSNR
  - 6.1% improvement in SSIM
  - Captures scene dynamics using voxel flow, while adjusting for camera trajectory
  - Localizes voxel flow and preserves the static scene
- Failure Cases:
  - Models other than M5 fail to capture object dynamics
  - The capture voxel flow is coarse, which may be attributed to the use of perceptual loss

Conclusions

- Our model disentangles real-world videos into a static scene, the camera trajectory and the scene flow that captures dynamics
- The flow architecture and robust losses enable learning
- Our model can generate plausible, novel middle frames

- Future work:
  - Use voxel flow to inform action classification
  - Use voxel flow for compression of videos with dynamic scenes
  - Experiment with more interpretable voxels or point clouds to better capture scene dynamics

References

[1] Z. Lai, S. Liu, A. A. Efros, and X. Wang, "Video autoencoder: self-supervised disentanglement of 3d structure and motion," in *ICCV*, 2021.
[2] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," CoRR, vol. abs/2011.13961, 2020.
[3] . Liu, R. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in ICCV, pp. 4473–4481, 10 2017.