Collision Prediction

Deep Learning (Fall 2023)

PRESENTED BY TEAM 27

- **Ritika**
- **Sujana**
- **Amardeep**

Problem Statement

- Given 11 frames of videos, where simple 3D shapes interact with each other according to basic physics principles, predict semantic segmentation of 22nd frame.

Our Approach

- Divided into two separate problems: Next Frame Prediction and Semantic Segmentation Prediction.
- **Next Frame Prediction**:
  - Started with Simvp, scope of improvisations were less.
  - Trained a diffusion based model to predict 22nd frame.
- **Semantic Segmentation:** U-Net

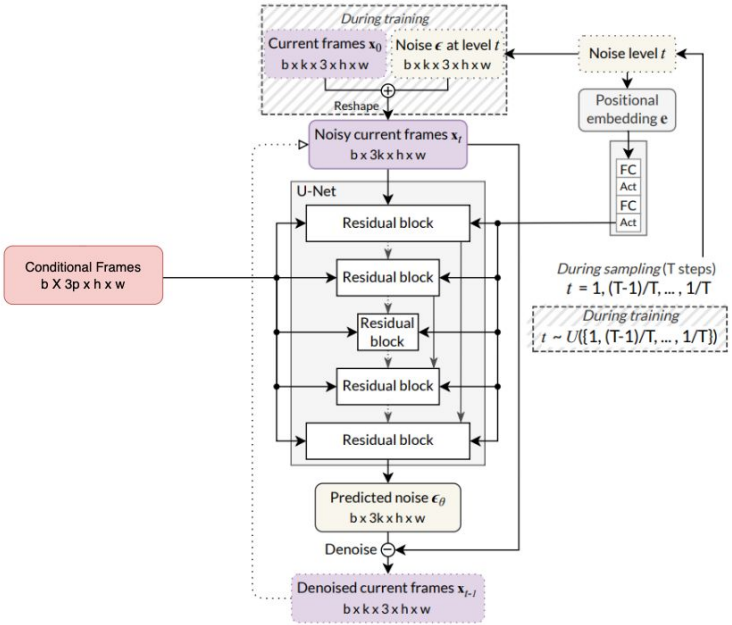
Next Frame Prediction via Conditional Diffusion

Conditioning a Diffusion Process

- Forward Diffusion Process:
  - $q_t(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}),$
- Reverse Diffusion Process:
  - $p_t(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}),$
- Loss functional for Network optimization:
  - $L(\theta) = \mathbb{E}_{t,\mathbf{x}_0\sim p_{\text{data}},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \mid t)\right\|_2^2\right]$
- But how to condition on past frames to generate current frames?
  - **p =** past frames[p], $\mathbf{x_0}$ = current frames[k]

$$L_{\text{vidpred}}(\theta) = \mathbb{E}_{t,[\mathbf{p},\mathbf{x}_0]\sim p_{\text{data}},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \mid \mathbf{p}, t)\right\|^2\right]$$

Network Architecture



Inspired from MCVD - Voleti et.al [NeurIPS - 2022]

NYU

Next Frame Prediction via Conditional Diffusion

Experiments and Results

#	Conditional frames	Future frames	Validation Jaccard Score (22^{nd} frame)	Observations
11v1(baseline)	11 given input frames	22^{nd} frame	19.8	• Loss and Jaccard score stagnated after 30 epochs on unlabeled data
11v1-AR	Sliding window of 11 frames	Next frame of the window (22^{nd} frame predicted auto regressively)	30.89	• Training data augmented by factor of 10x due to sliding window • Inference time increased by 11x due to autoregressive prediction of 22^{nd} frame. • Model started overfitting after 40 epochs on unlabeled data, with MSE of 0.004 between target and predicted frame. • 1000 subsampling steps gave best results
5V6	Alternate 5 frames(2, 4, 6, 8, 10)	Alternate 6 frames (12, 14, 16, 18, 20, 22)	13.6	• Neither noise loss converged nor MSE between target and predicted frames.

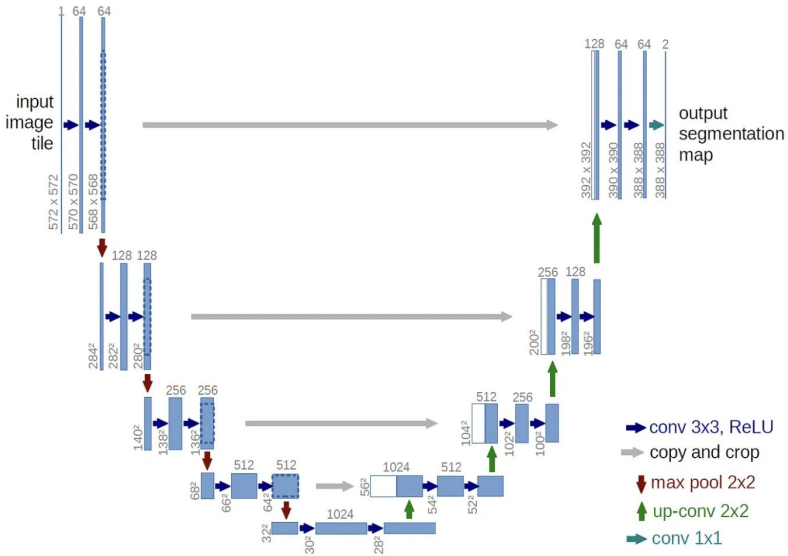
U-Net for Segmentation of Predicted Frame

U-Net Architecture

- 4 Encoder Layers containing convolutions followed by max-pooling for downsampling
- 1 Bottleneck Layer
- 4 Decoder Layers containing upconvolutions for upsampling
- Skip Connections, each encoder layer is connected to its corresponding decoder layer

Epochs	Validation Jaccard
15	0.966
30	0.969

Unet Diagram



Future work and improvements

- Using conditional diffusion to predict semantic segmentation frame, instead of raw video frames.
- Make sub-sampling process of diffusion process efficient.
- Train better Semantic segmentation model.