Towards Scalable Neural Representation for Diverse Videos

Bo He[1], Xitong Yang[2], Hanyu Wang[1], Zuxuan Wu[3], Hao Chen[1], Shuaiyi Huang[1], Yixuan Ren[1], Ser-nam Lim[2], Abhinav Shrivastava[1]

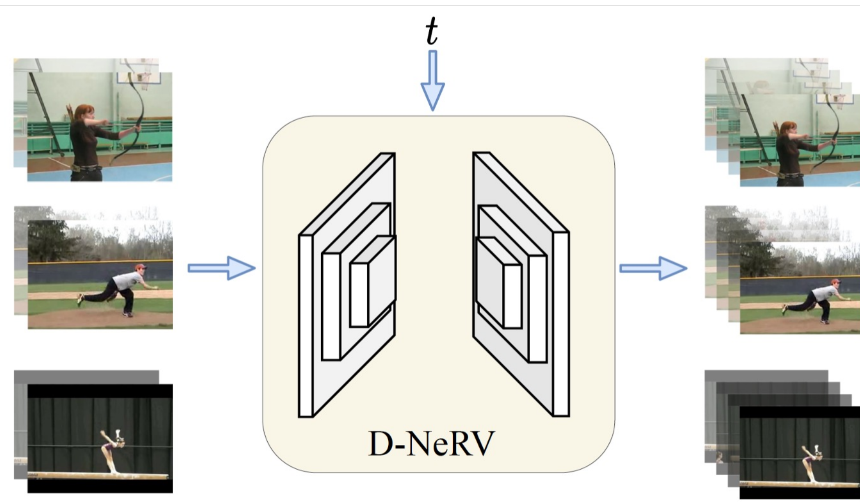
[1] University of Maryland, College Park [2] Meta AI [3] Fudan University

Poster: TUE-PM-192

Website and Code: https://boheumd.github.io/D-NeRV

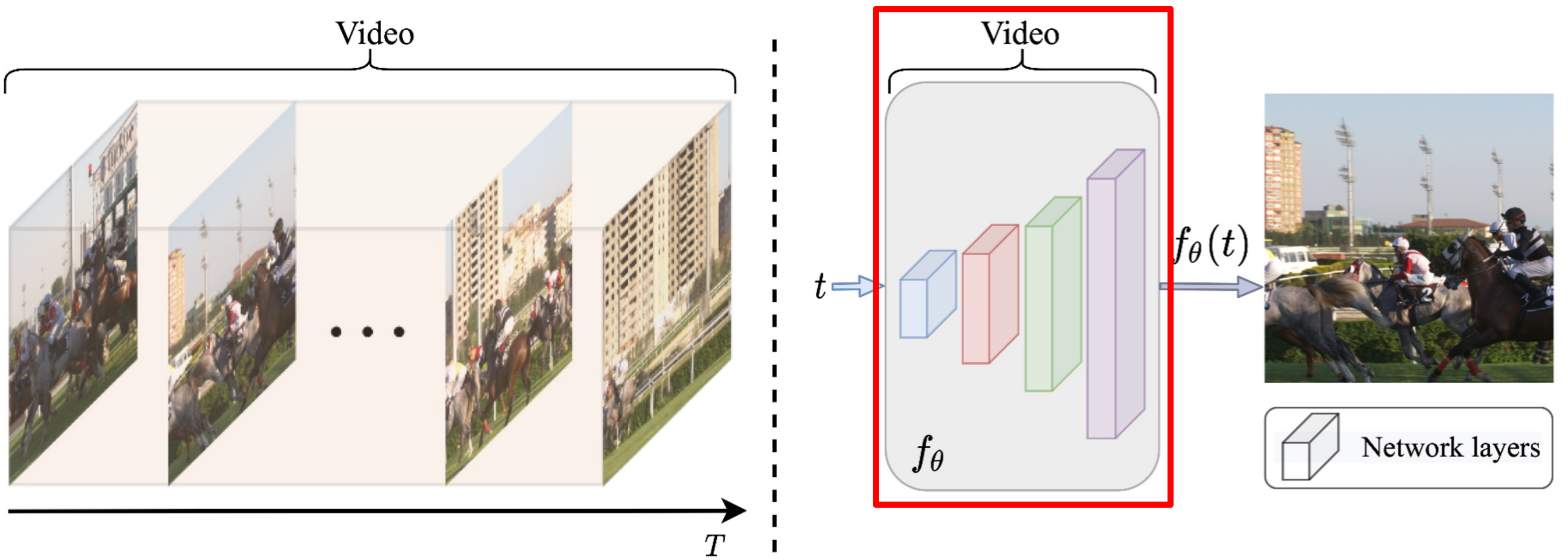
Overview

- D-NeRV is an implicit neural representation designed for **large-scale and diverse** videos.
- D-NeRV achieves SoTA on the **video compression** task.
- D-NeRV shows its advantages as an efficient and effective dataloader for downstream **video understanding** task.



Background

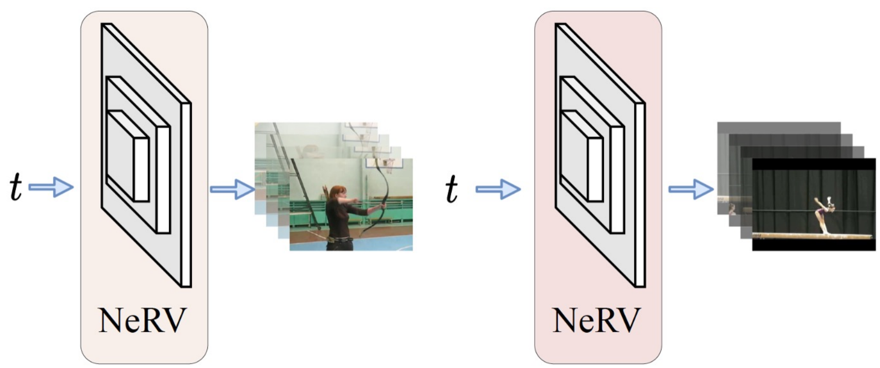
- Video Representations: Explicit vs. Implicit

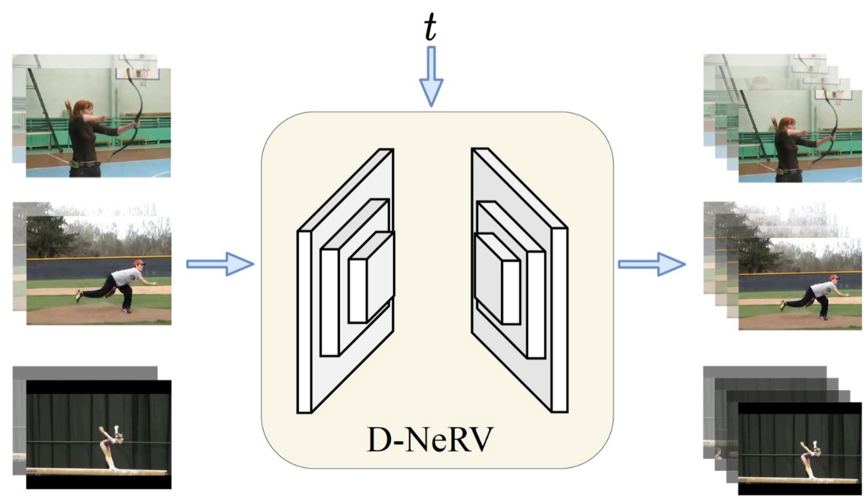


(a) Explicit representations for videos (e.g., H265) (b) Neural implicit representations for videos (e.g., NeRV)

Chen et al. *"NeRV: Neural Representations for Videos."* In NeurIPS, 2021.

Comparison of NeRV and D-NeRV





NeRV

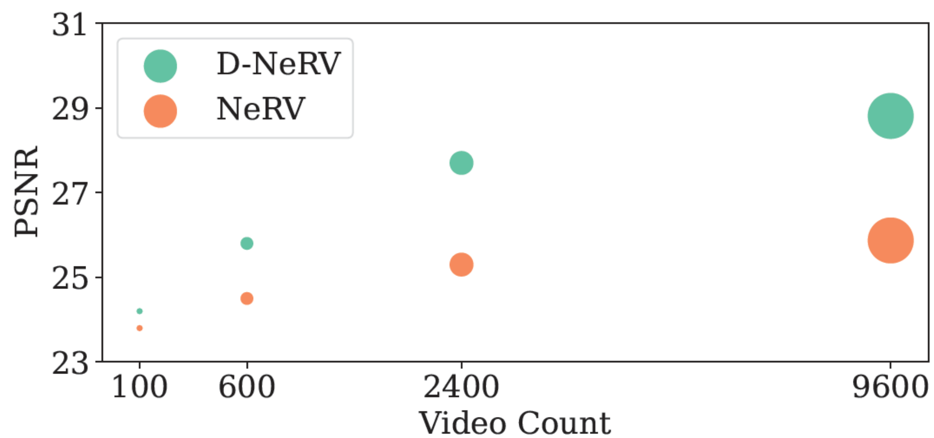
- Limited to encode several **short** videos.
- Optimize representation to each video *independently*.

D-NeRV

- Designed for encoding **large-scale and diverse** videos.
- Encodes all videos into a *shared* model by conditioning on *keyframes*.

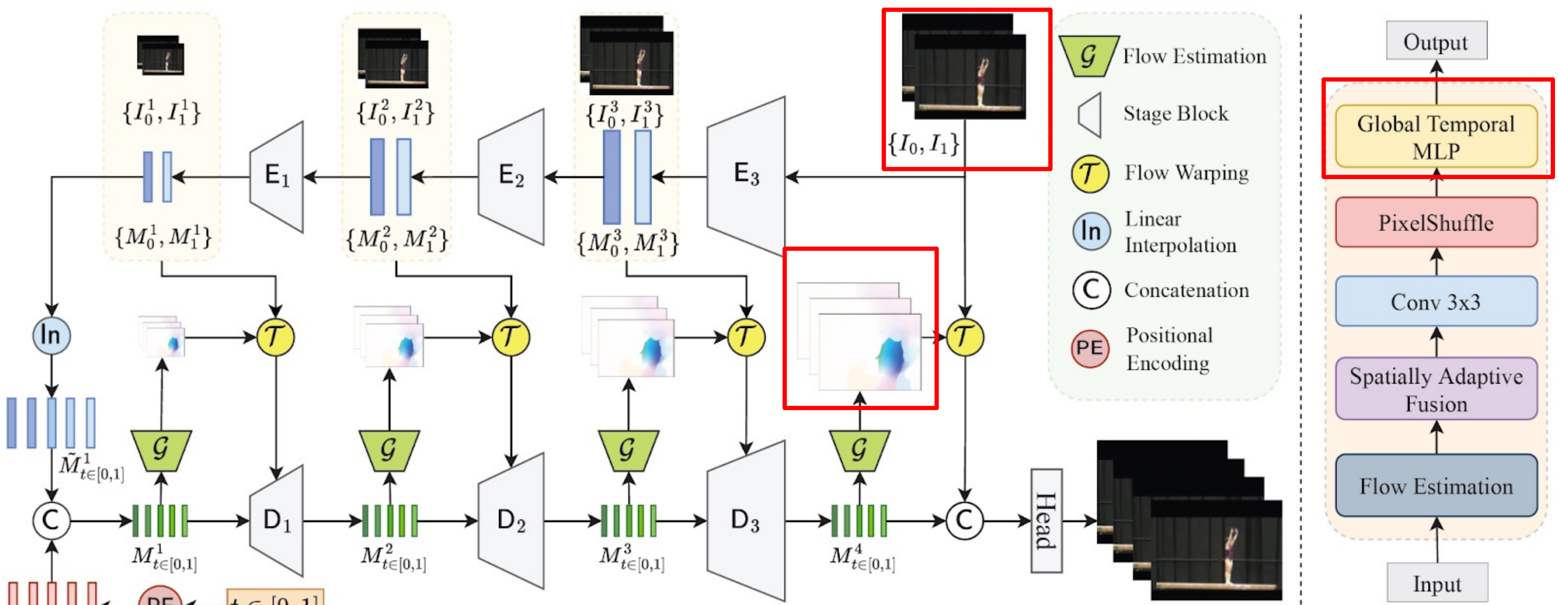
Motivation

- **Motivation:** How to encode **large-scale and diverse** videos using the implicit neural representation?
- **Naive Solution:** Encode each video with a ***separate*** model.
- **Observation:** Encode diverse videos in a ***shared*** model, PSNR performance increases as the video count increases.



Architecture

- **Encoder-Decoder:** conditions on keyframes from each video
- **Optical Flow:** reduces spatial redundancies
- **Temporal Modeling:** model relationship across frames

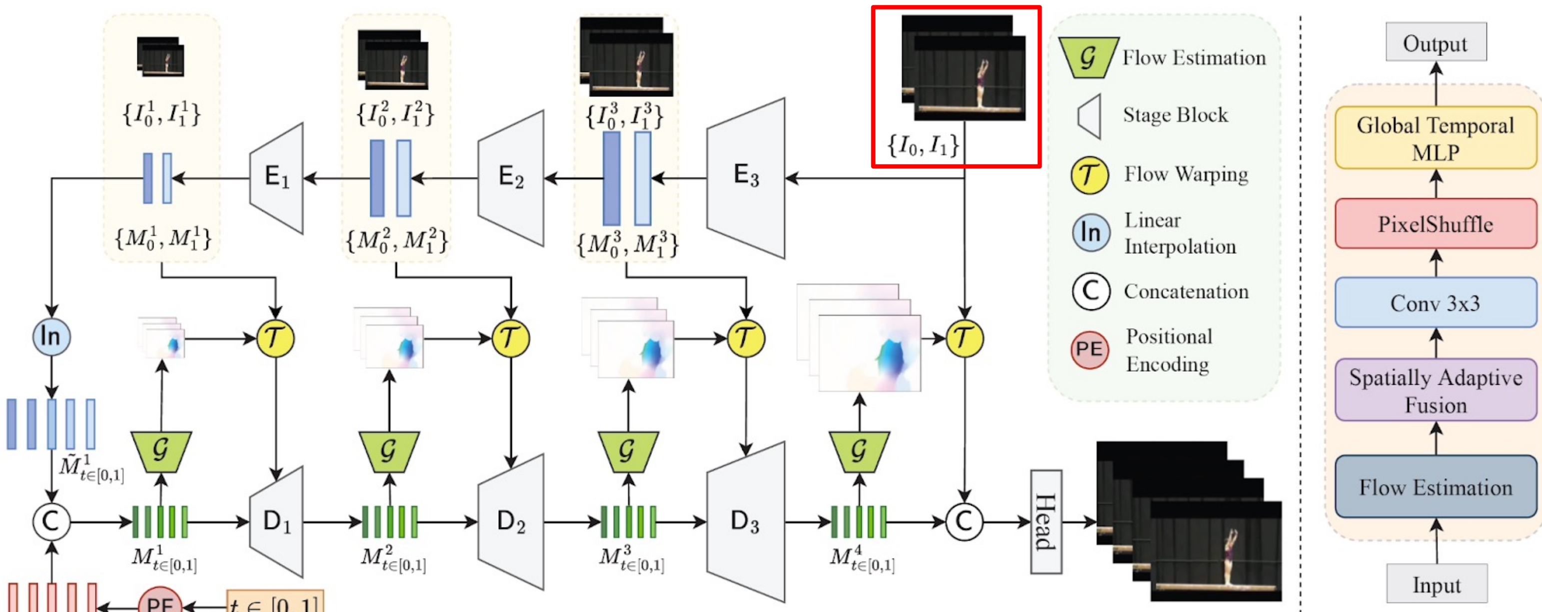


(a) Framework Overview
(b) Decoder Block

Visual Content Encoder

Motivation: Visual content of each video varies significantly, directly memorizing all videos introduces too much optimization complexity.

Solution: Sample keyframes from each video as conditional input.

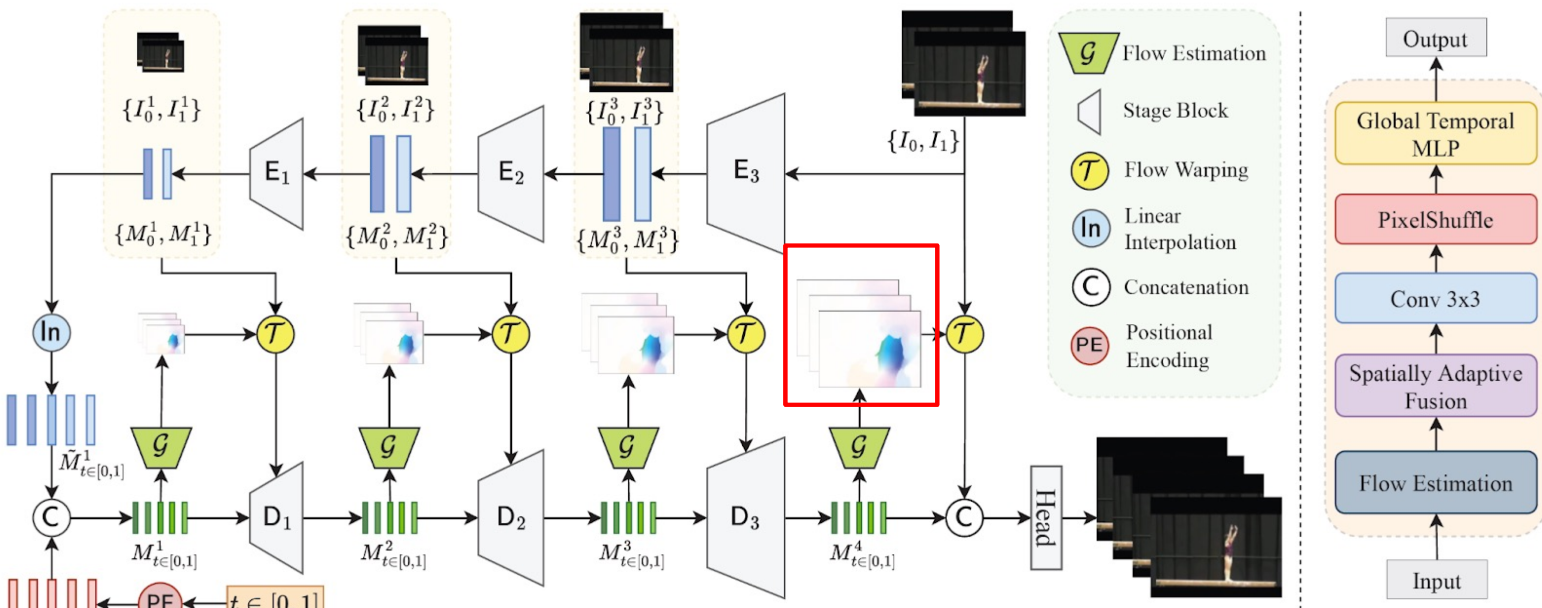


(a) Framework Overview (b) Decoder Block

Task-oriented Optical Flow

Motivation: Reduce spatial redundancies across frames in the RGB space.

Solution: Predict task-oriented optical flow w.r.t. the keyframes.

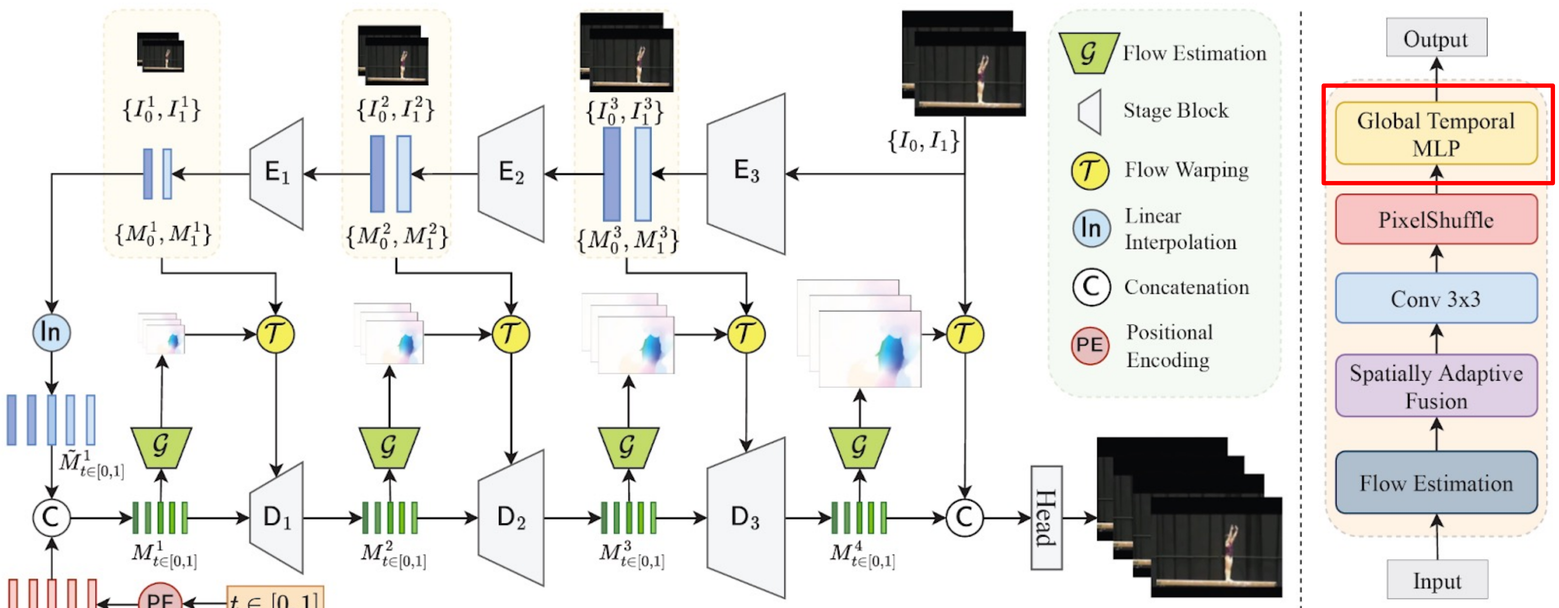


(a) Framework Overview (b) Decoder Block

Global Temporal Modeling

Motivation: NeRV outputs each frame independently and neglects temporal relationship across frames.

Solution: Model global temporal relationship for each video clip.



(a) Framework Overview (b) Decoder Block

Ablation Studies

Model	UVG		UCF101	
	PSNR	MS-SSIM	PSNR	MS-SSIM
NeRV	34.13	0.948	28.00	0.935
+ GTMLP	33.94	0.946	27.96	0.935
+ SAF	35.84	0.960	30.78	0.962
+ GTMLP	36.32	0.963	30.94	0.964
+ Flow	**36.99**	**0.977**	**31.44**	**0.968**

Contribution of each components

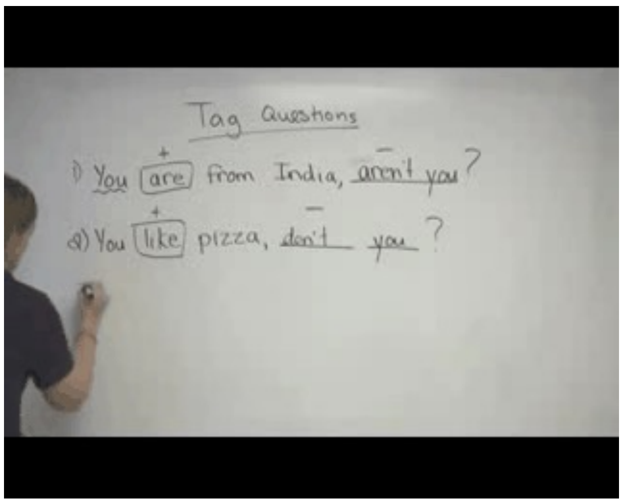
- SAF: spatially-adaptive fusion
- GTMLP: global temporal MLP
- Flow: multi-scale flow estimation

Video Diversity Ablation

1. Total video count fixed to 1000.
2. Change the number of action classes (diversity).
3. D-NeRV is more capable of representing diverse videos.

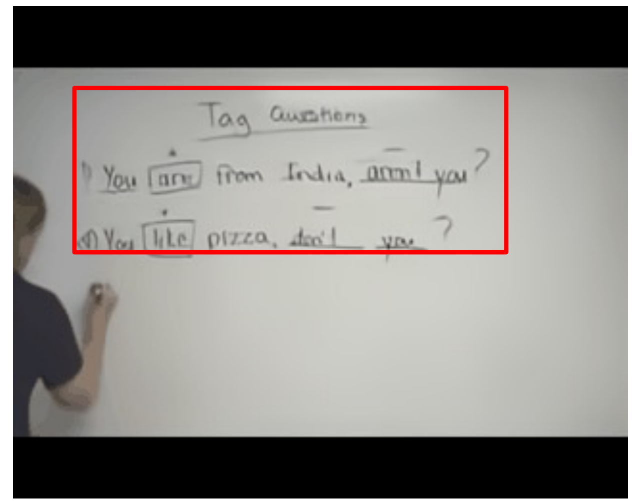
	#Class	PSNR	MS-SSIM
NeRV	10	27.95	0.935
	100	26.66	0.915
	$\triangledown$	-1.29	-0.02
D-NeRV	10	29.74	0.950
	100	29.36	0.946
	$\triangledown$	**-0.38**	**-0.004**

Video Compression Visualization



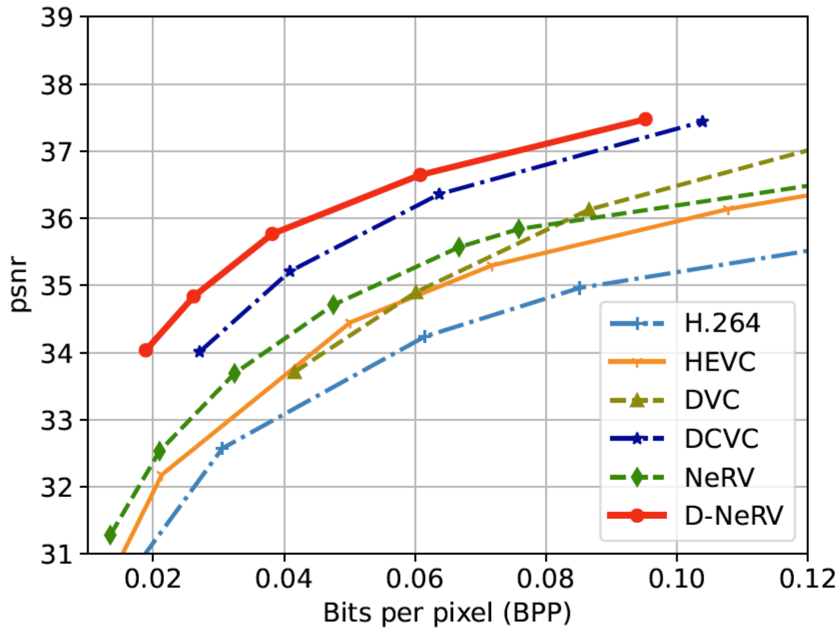
Ground Truth

NeRV

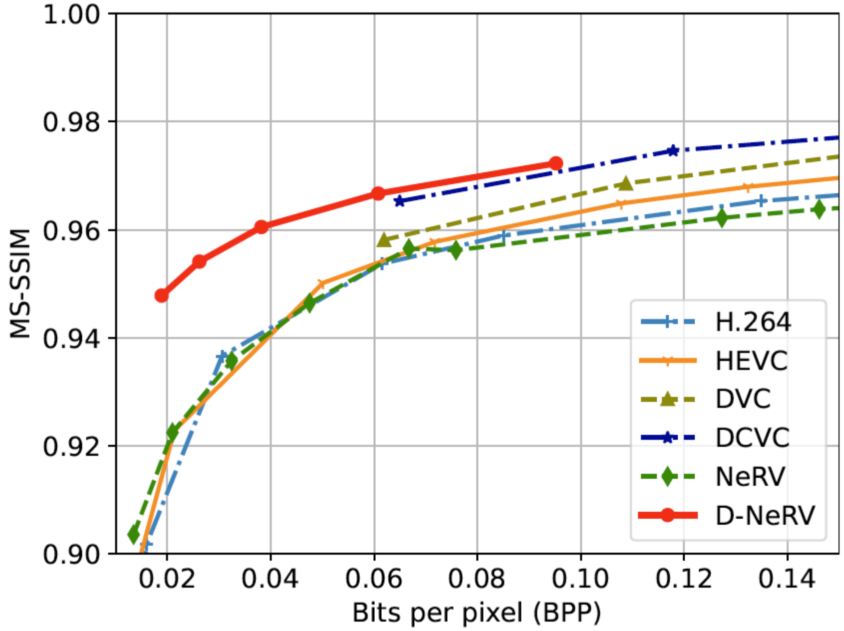


D-NeRV

Video Compression



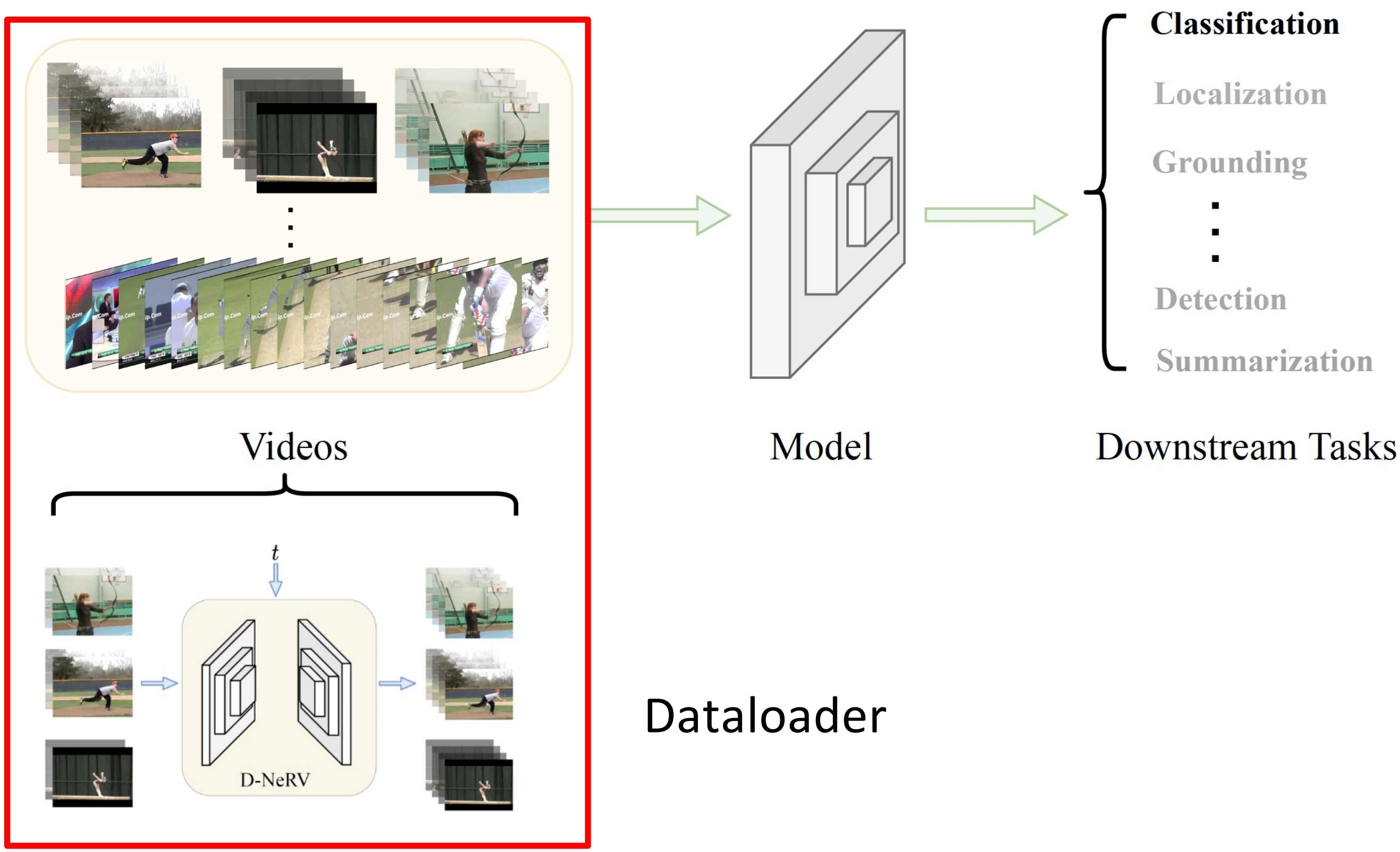
(a) PSNR vs. BPP



(b) MS-SSIM vs. BPP

UVG dataset

D-NeRV as Dataloader



Downstream Action Recognition Task

- D-NeRV encodes the whole UCF101 dataset, use it as dataloader

Model	S	M	L
GT	91.3	91.3	91.3
H.264	77.2	82.4	85.5
NeRV	71.9	75.9	80
D-NeRV	**81.1**	**84.4**	**86.4**

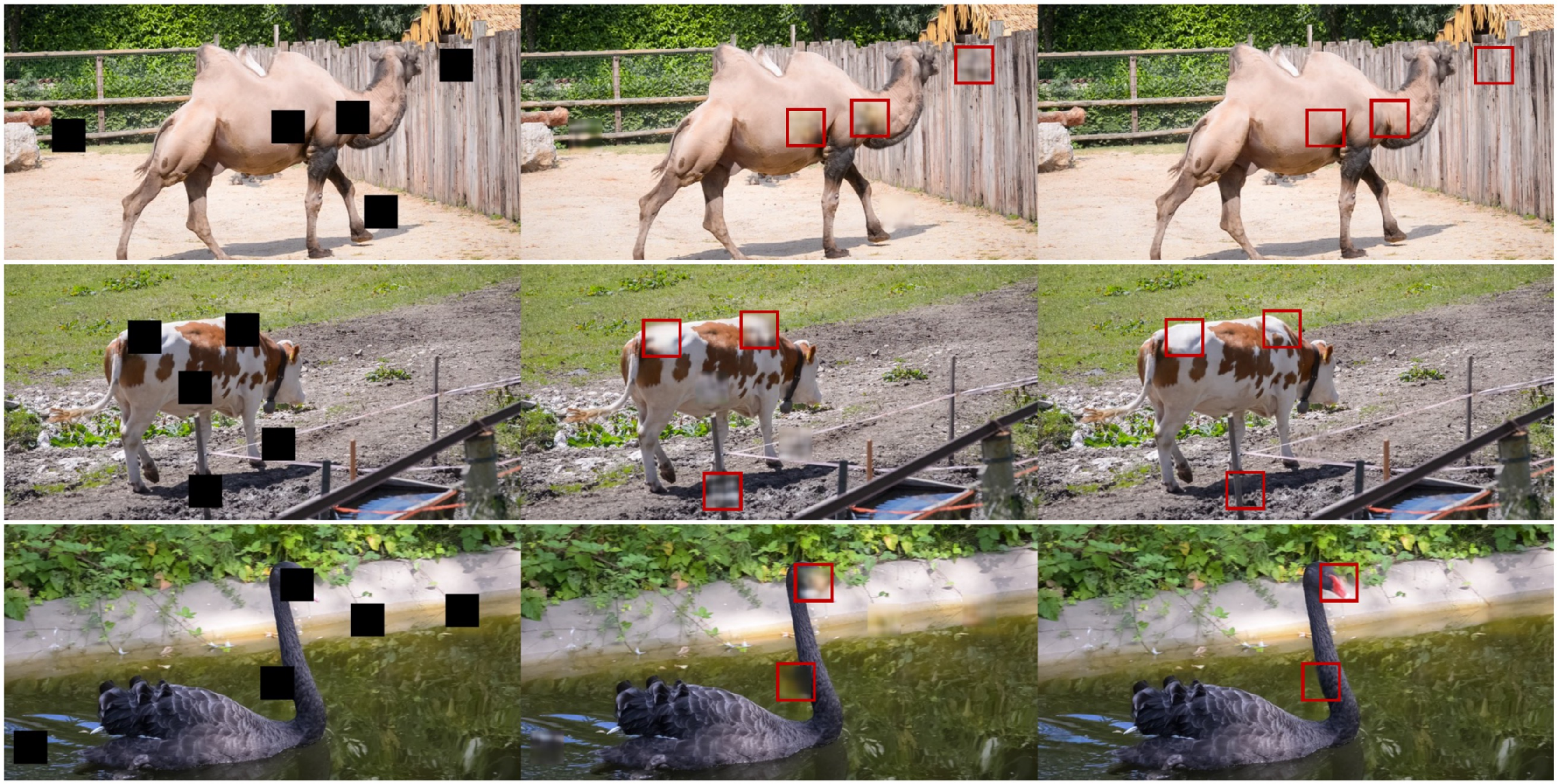
Action recognition

Method	VPS ↑
Frame (Tab. 8 GT)	273
H.264	265
DCVC	0.9
NeRV (fp32)	383
D-NeRV (fp32)	266
NeRV (fp16)	454
D-NeRV (fp16)	363

Decoding speed

Decoding speed is much faster than learning-based methods

Video Inpainting Visualization



(a) Ground Truth (b) NeRV (c) D-NeRV

Thank You!

For more details, please visit

Poster# 192 at

20-Jun-23, 4:30pm-6:30pm

Website and Code:

https://boheumd.github.io/D-NeRV