Scalable Neural Video Representations with Learnable Positional Features

Subin Kim[*,1], Sihyun Yu[*,1], Jaeho Lee[2], Jinwoo Shin[1]

[1]Korea Advanced Institute of Science and Technology (KAIST)
[2]Pohang University of Science and Technology (POSTECH)

Project page | Paper | Code

TL;DR: We propose a compute-/memory-efficient neural representation for videos

Summary

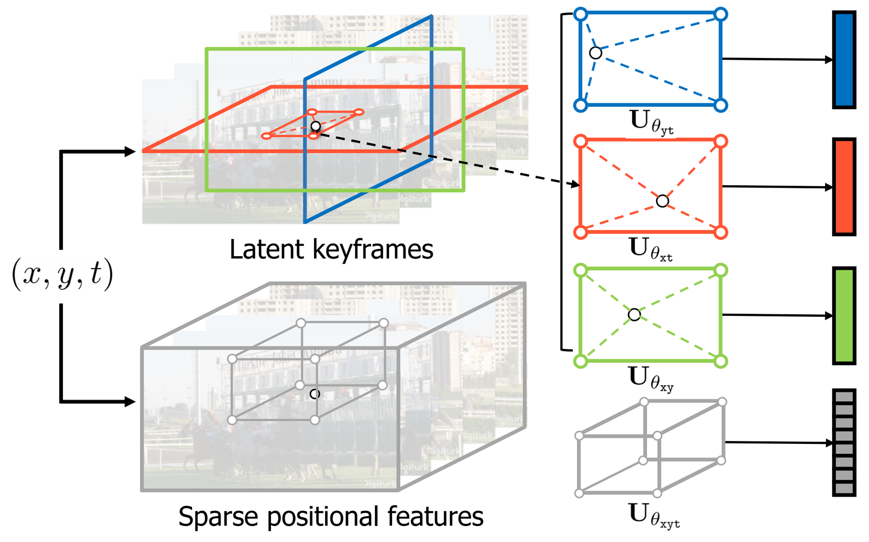
NVP can capture the detail of a video containing dynamic motions after training for "**1 minute**".


NeRV(BPP: 0.938) Instant-ngp (BPP: 6.489) NVP (ours, BPP: 0.189) Ground Truth

Motivation: Recent advances in coordinate-based neural representations (CNRs) have shown great promise in the field as a new paradigm for representing complex signals. However, video CNRs often suffer from two inefficiencies that prevent them from practical usage; (1) severe compute-inefficiency and (2) sacrifice of the parameter-efficiency.

Contribution: We introduce a *neural video representation with learnable positional features* (**NVP**), a novel CNR for videos that is the best of both worlds, **achieving high-quality encoding** and the **compute-/parameter-efficiency simultaneously**.

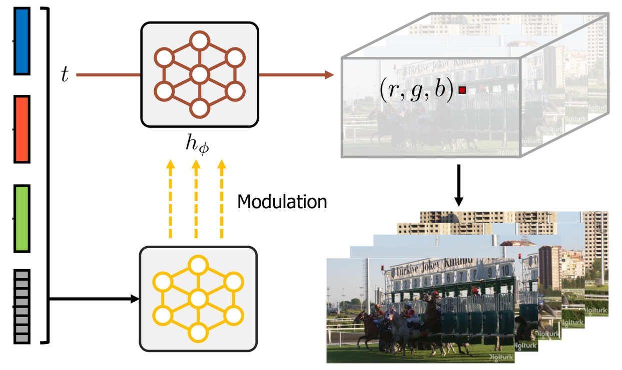
Amortize a Given Video as Succinct Latent Grids



Learnable latent keyframes ($U_{\theta_{xy}}, U_{\theta_{xt}}, U_{\theta_{yt}}$)
- **Image-like 2D** latent grids **across t-, y-, x-axis**.
- **Learned** to capture **common representative contents** over spatio-temporal directions.

Sparse positional features ($U_{\theta_{xyt}}$)
- **Video-like 3D** sparse latent grids.
- Despite its sparsity, it **efficiently captures the local details of the video** since the common contents of a given video are effectively encoded with the latent keyframes.

Modulate the Latent Codes to the RGB Values



Modulated implicit function (h_ϕ)
- Maps a latent vector to the corresponding RGB value.
- Design h_ϕ to be a K-layer Multi-layer perceptron (MLP) modulated by another modulator network, instead of simple MLP (more expressive power).

Compute-/Memory-efficient Compression Procedure

Incorporate powerful existing image & video codecs to compress our latent features
- Quantize latent keyframes and sparse positional features as 2D/3D grids of 8-bit latent codes.
- Regard the quantized latent codes as image and video pixels and compress them using codecs.
- Notably **maintaining the video quality without any fine-tuning** (compute-efficient).

Quantitative Results

Compute-efficiency; achieves reasonable encoding quality within a short training cost.

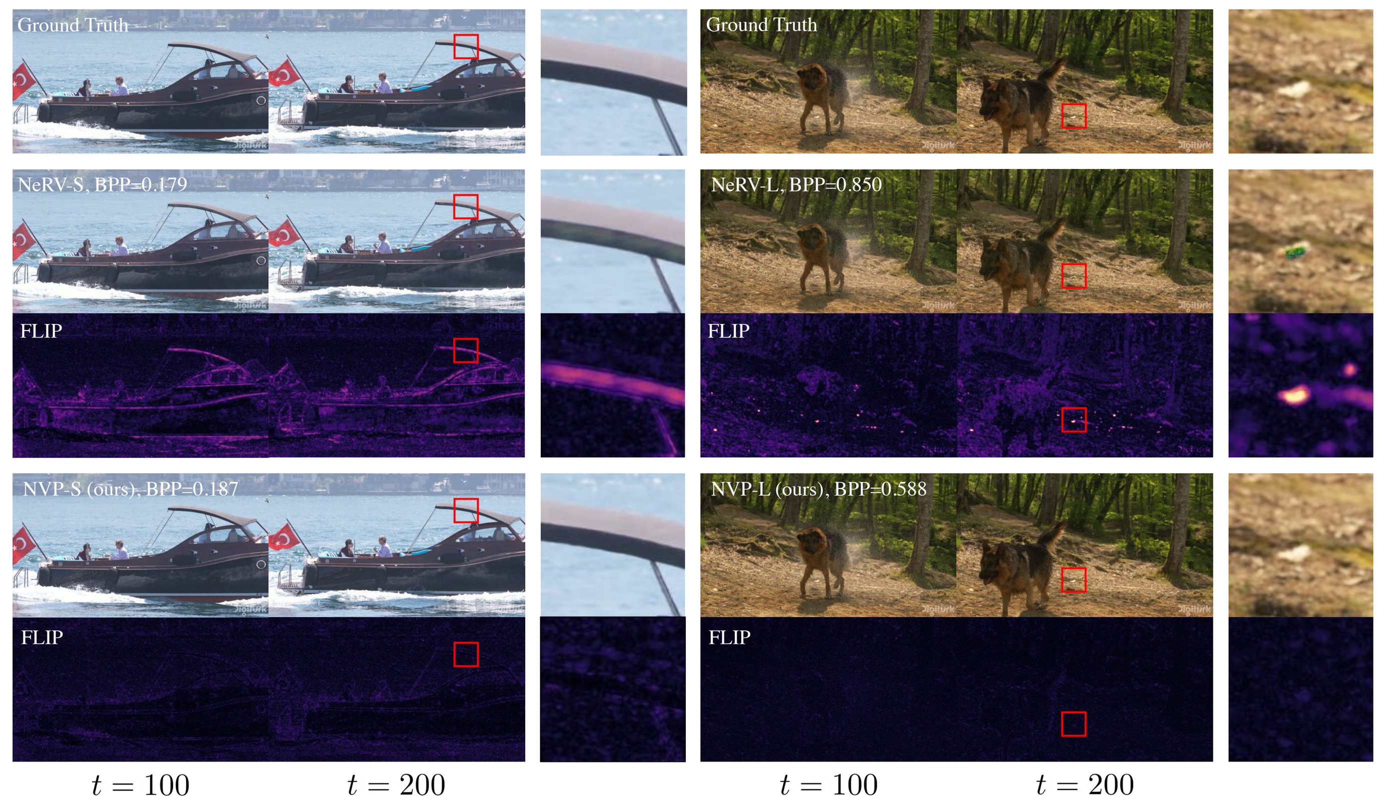
Encoding time	Method	BPP	PSNR (↑)	FLIP (↓)	LPIPS (↓)
~5 minutes	Instant-ngp [34]	7.580	33.15±3.19	0.090±0.034	0.226±0.112
	NeRV-S* [5]	1.072	24.16±5.17	0.219±0.097	0.542±0.180
	NVP-S* (ours)	0.901	**34.57±2.62**	**0.075±0.021**	**0.190±0.100**
~10 minutes	Instant-ngp [34]	7.580	34.07±3.01	0.082±0.030	0.204±0.105
	NeRV-S* [5]	1.072	26.53±5.92	0.176±0.088	0.460±0.184
	NVP-S* (ours)	0.901	**35.79±2.31**	**0.065±0.016**	**0.160±0.098**
~1 hour	Instant-ngp [34]	7.580	35.69±2.72	0.071±0.025	0.162±0.090
	NeRV-S* [5]	1.072	33.26±4.31	0.094±0.038	0.240±0.112
	NVP-S* (ours)	0.901	**37.61±2.20**	**0.052±0.011**	**0.145±0.106**

Parameter-efficiency; succinct neural representation with a high-quality encoding.

~15 hours	SIREN [40]	0.284	27.20±3.77	0.169±0.059	0.409±0.124
	FFN [46]	0.284	28.18±3.62	0.153±0.055	0.442±0.126
	Instant-ngp [34]	0.229	28.81±3.48	0.155±0.057	0.390±0.135
	NeRV-S [5]	0.201	36.14±3.97	0.067±0.023	0.163±0.101
~8 hours	**NVP-S (ours)**	0.210	**36.46±2.18**	**0.067±0.017**	**0.135±0.083**
>40 hours	SIREN [40]	0.284	26.09±3.88	0.175±0.082	0.486±0.188
	FFN [46]	0.284	29.53±3.44	0.135±0.052	0.391±0.124
	Instant-ngp [34]	0.436	29.98±3.39	0.138±0.051	0.358±0.140
	NeRV-L [5]	0.485	35.00±3.31	0.079±0.020	0.145±0.100
~11 hours	**NVP-L (ours)**	0.412	**37.47±2.08**	**0.062±0.017**	**0.102±0.061**

Qualitative Results

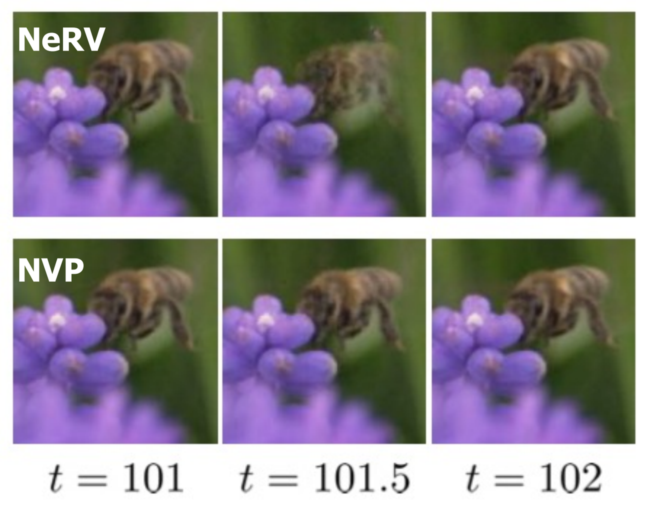
NVP does not suffer from undesirable artifacts when compressed.



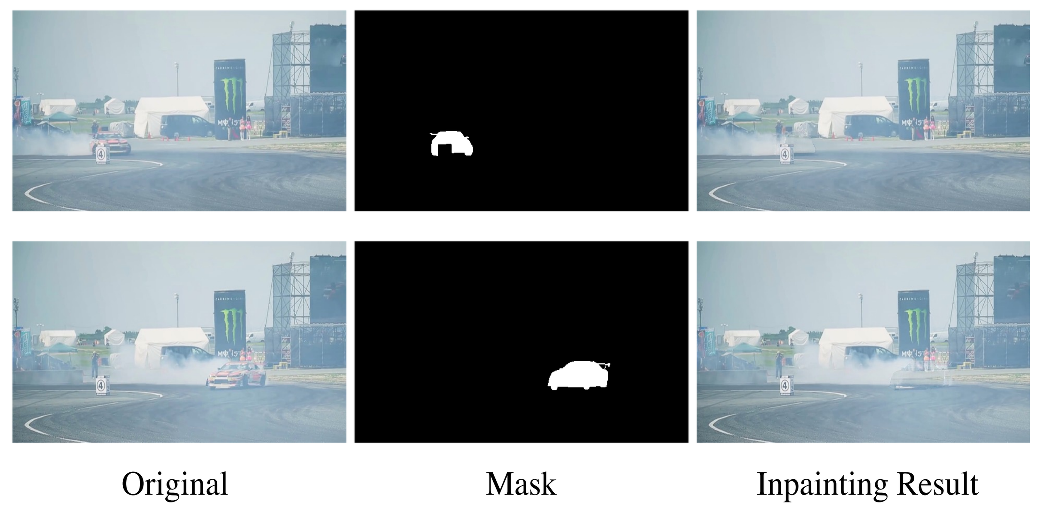
Various Application of NVP as a Video CNR

NVP can show numerous compelling properties as a video CNR.

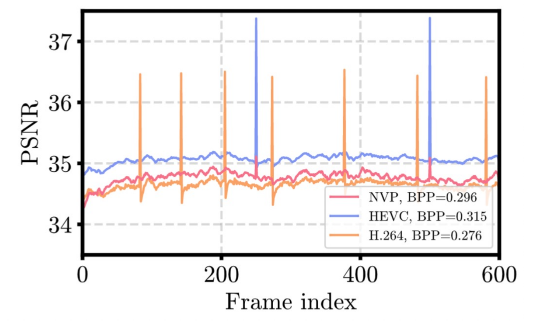
Video Frame Interpolation.



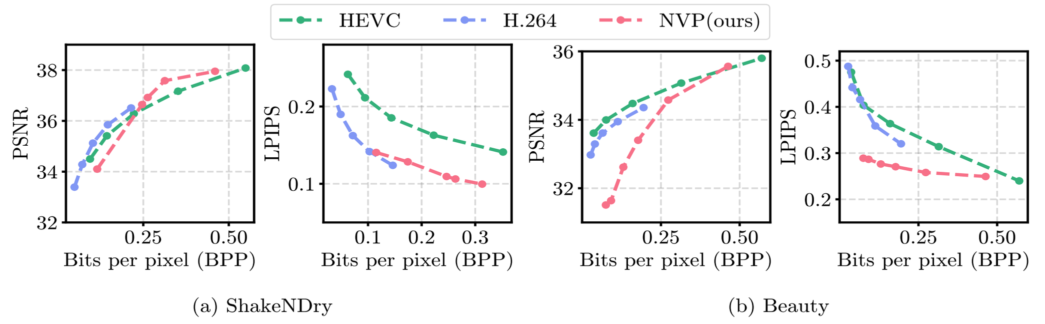
Video Inpainting.



Consistent frame-wise encoding.



Video Compression.



See the paper for more experiments, including ablation studies and detailed explanations.
For better, playable illustrations and qualitative results, please refer to our project page. ☺