Boosting Neural Representations for Videos with a Conditional Decoder

Xinjie Zhang, Ren Yang, Dailan He, Xingtong Ge, Tongda Xu, Yan Wang, Hongwei Qin, Jun Zhang

CVPR 2024 Highlight

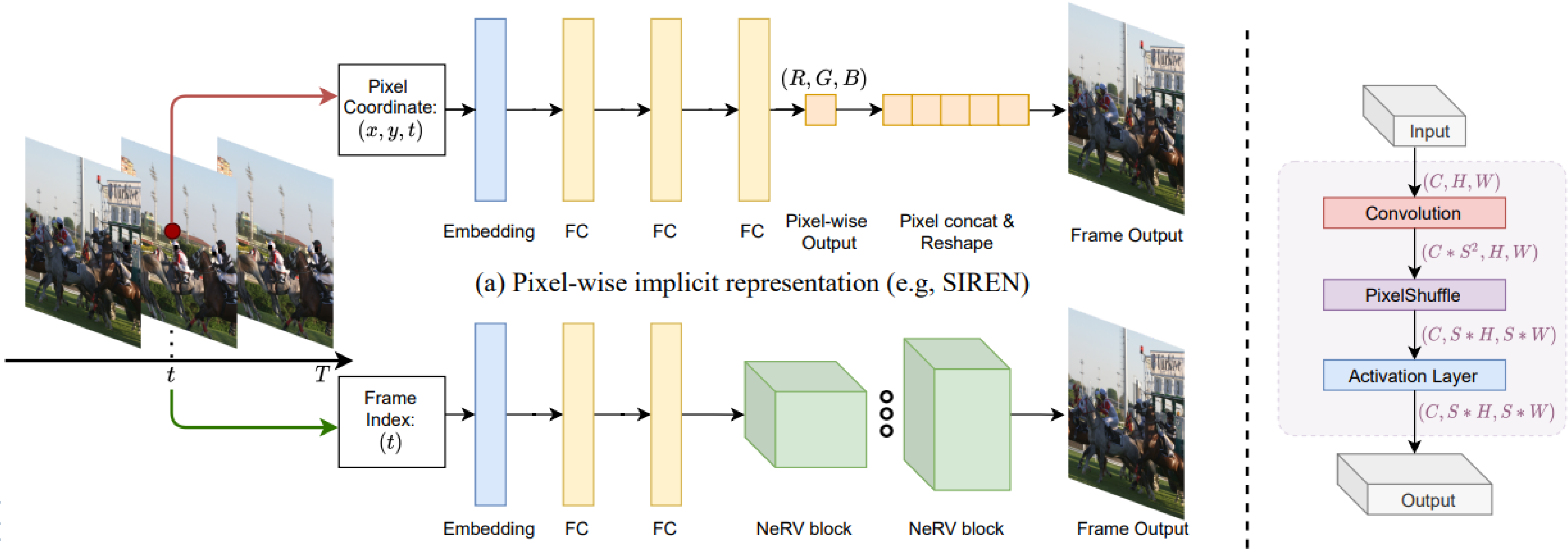
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Background: Neural Representation for Video

NeRV: pioneer of image-wise implicit video representation

- Higher fitting quality
- Faster decoding speed
- Fewer training time

Methods	Parameters	Training Speed ↑	Encoding Time ↓	PSNR ↑	Decoding FPS ↑
SIREN [5]	3.2M	1×	2.5×	31.39	1.4
NeRF [4]	3.2M	1×	2.5×	33.31	1.4
NeRV-S (ours)	3.2M	**25**×	**1**×	**34.21**	**54.5**
SIREN [5]	6.4M	1×	5×	31.37	0.8
NeRF [4]	6.4M	1×	5×	35.17	0.8
NeRV-M (ours)	6.3M	**50**×	**1**×	**38.14**	**53.8**
SIREN [5]	12.7M	1×	7×	25.06	0.4
NeRF [4]	12.7M	1×	7×	37.94	0.4
NeRV-L (ours)	12.5M	**70**×	**1**×	**41.29**	**52.9**



(a) Pixel-wise implicit representation (e.g, SIREN)

(b) NeRV: Image-wise implicit representation

(c) NeRV block

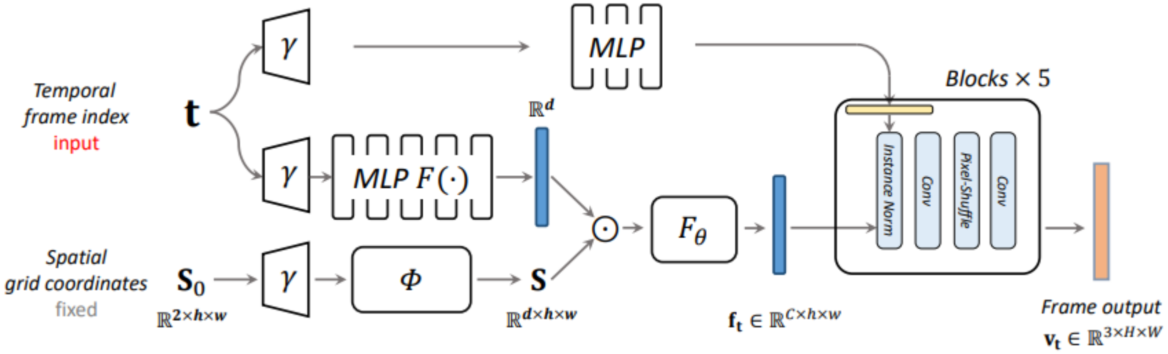
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

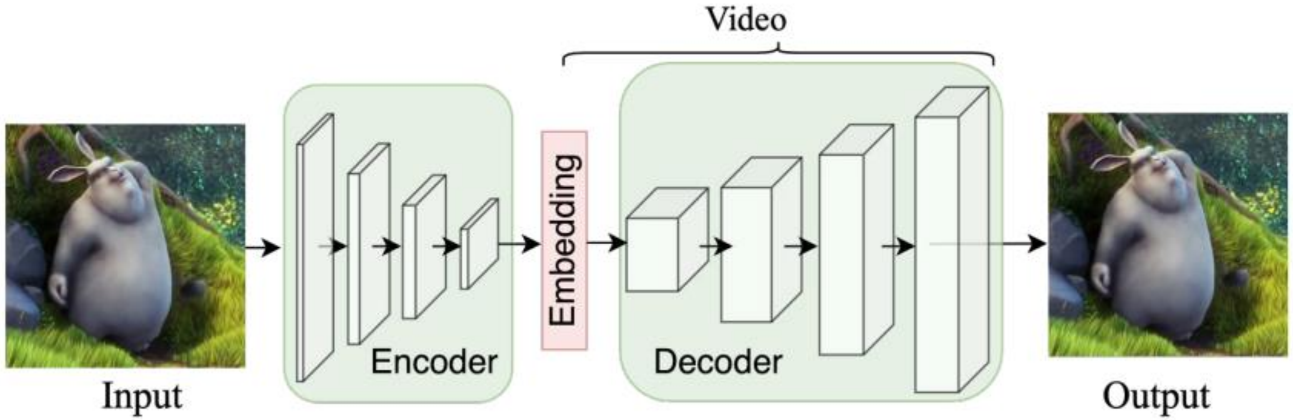
[1] Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser Nam Lim, and Abhinav Shrivastava. Nerv: Neural representations for videos. NeurIPS, 2021.

Background: Neural Representation for Video

A series of subsequent works design more meaningful embeddings to improve the quality of video reconstruction.

- E-NeRV [ECCV'22]: Spatial-temporal positional embedding
- HNeRV [CVPR'23]: Content-aware embedding
- …



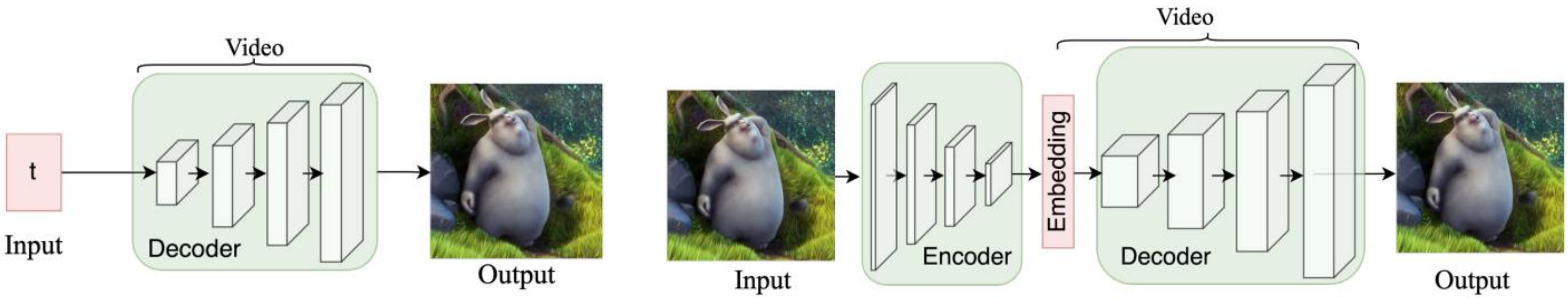


[2] Zizhang Li, Mengmeng Wang, Huaijin Pi, Kechun Xu, Jianbiao Mei, and Yong Liu. E-nerv: Expedite neural video representation with disentangled spatial-temporal context. ECCV, 2022.
[3] Hao Chen, Matthew Gwilliam, Ser-Nam Lim, Abhinav Shrivastava. Hnerv: A hybrid neural representation for videos. CVPR, 2023.

Challenges: Representation

When decoding the *t*-th frame,

- Most works (NeRV, HNeRV, …) only relies on the *t*-th temporal embedding.
  - struggle to align the intermediate features with the target frame.



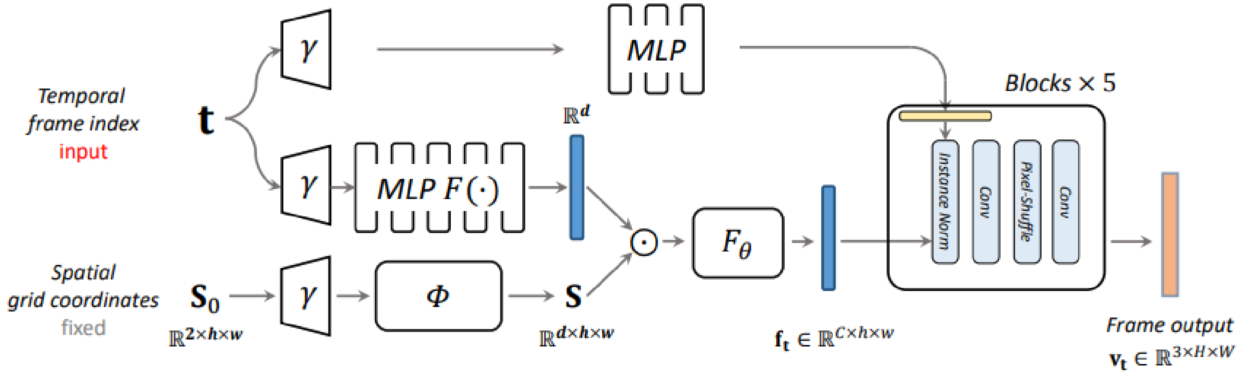
Challenges: Representation

When decoding the *t*-th frame,

- Most works (NeRV, HNeRV, …) only relies on the *t*-th temporal embedding.
  - struggle to align the intermediate features with the target frame.

- Few works (E-NeRV, …) adopt AdaIN module [4] to modulate intermediate features.
  - normalization operation might reduce the over-fitting capability of INR, resulting in limited performance gains.

$$(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t) = \mathrm{MLP}(\boldsymbol{z}_t),$$
$$\mathrm{AdaIN}(\boldsymbol{f}_t|\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t) = \boldsymbol{\sigma}_t \left(\frac{\boldsymbol{f}_t - \boldsymbol{\mu}(\boldsymbol{f}_t)}{\boldsymbol{\sigma}(\boldsymbol{f}_t)}\right) + \boldsymbol{\mu}_t,$$

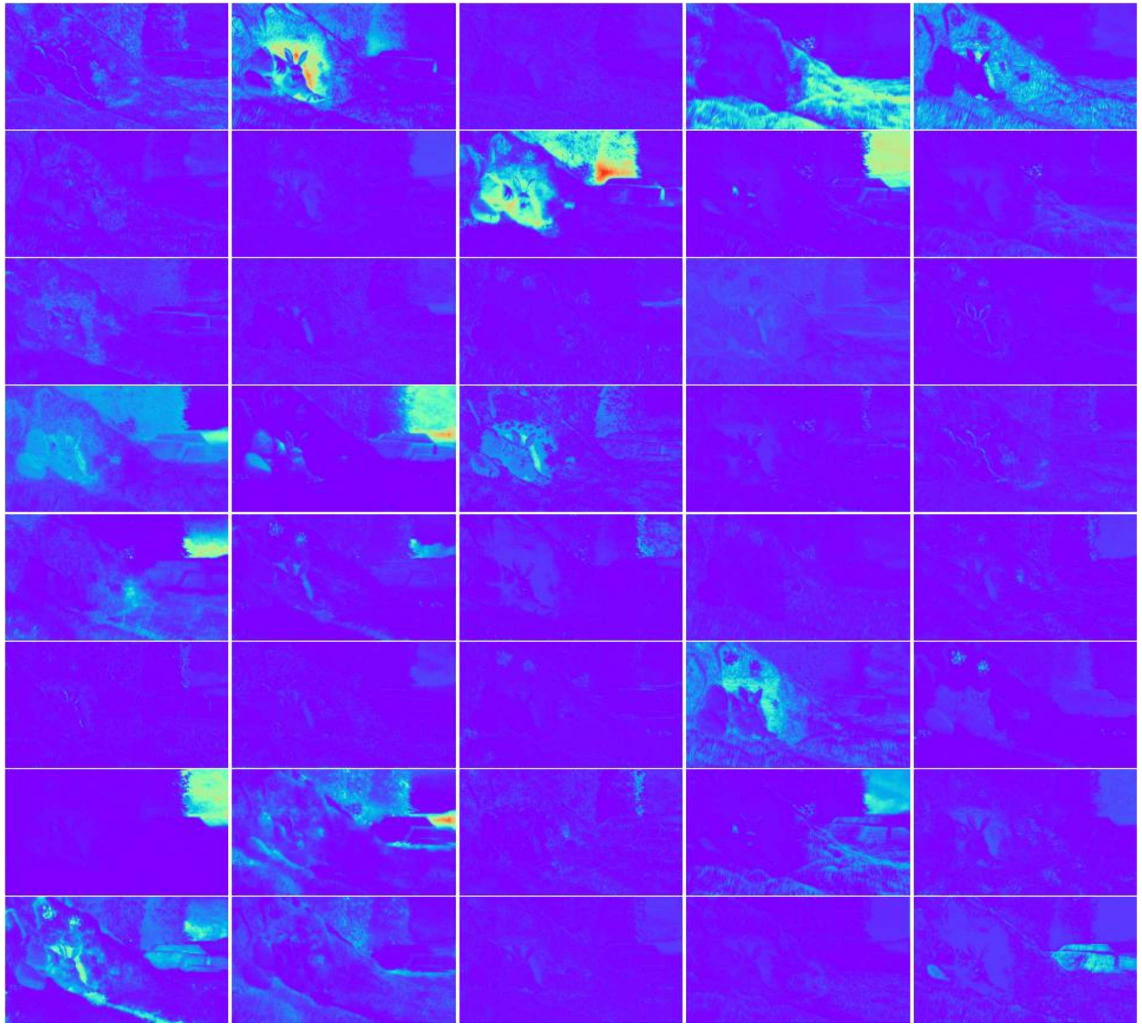


[4] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. CVPR, 2017.

Challenges: Representation

When designing the specific up-sampling block,

- ➢ Existing studies: refine NeRV's upsampling block for a more streamlined convolutional framework.
- ➢ Activation layers: the impact on the model's representational ability remains under-explored.
- ➢ GELU function: activate only a limited number of feature maps.



THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Challenges: Representation

When designing the loss function,

- Previous works rely on L2 loss or a combination of L1 and SSIM losses.
- Fail to preserve high-frequency information (e.g., edges and fine details within each frame)

How to improve the efficiency of NeRV methods?

Temporal-aware Affine Transform!

Sinusoidal NeRV-like Block!

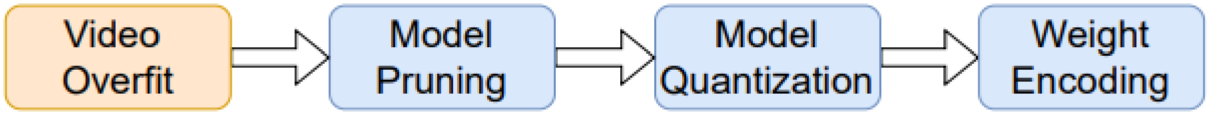
High-frequency Supervision!

Challenges: Compression

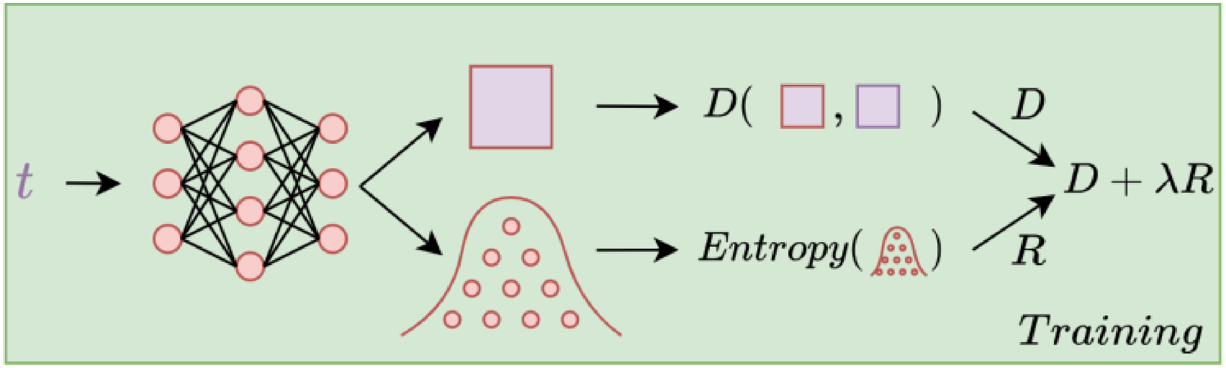
In video compression, NeRV methods covert the video compression problem to the model compression task.

- PQE: these components are optimized separately [1,2,3].
- Entropy minimization: inconsistency in the entropy models employed during both training and inference stages [5,6].

Consistent Entropy Minimization!



PQE: Three-step model compression pipeline



Entropy minimization: joint optimization of quantization and entropy coding

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

[5] Carlos Gomes, Roberto Azevedo, and Christopher Schroers. Video compression with entropy-constrained neural representations. CVPR 2023.
[6] Shishira R Maiya, Sharath Girish, Max Ehrlich, Hanyu Wang, Kwot Sin Lee, Patrick Poirson, Pengxiang Wu, Chen Wang, and Abhinav Shrivastava. Nirvana: Neural implicit representations of videos with adaptive networks and autoregressive patch-wise modeling. CVPR 2023.

Method: Universal Boosting-NeRV Framework

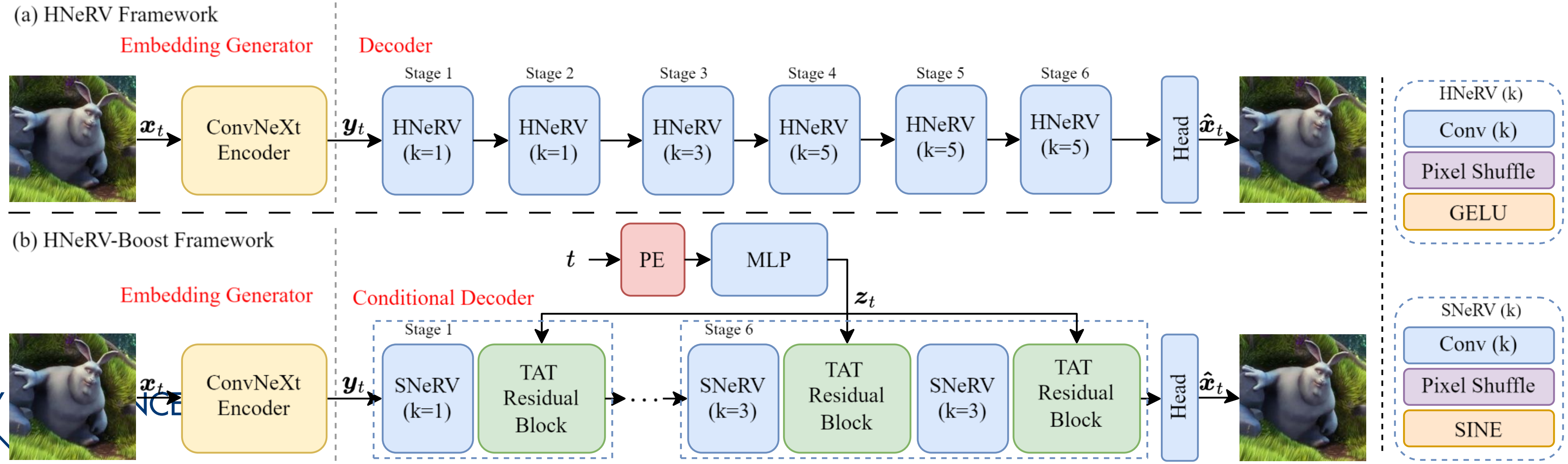
Two primary components:

- an embedding generator (→ specific video INR model)
- a conditional decoder

Boosted HNeRV:

$$\boldsymbol{y}_t = E(\boldsymbol{x}_t; \boldsymbol{\phi}),$$
$$\boldsymbol{z}_t = M(\mathrm{PE}(t); \boldsymbol{\psi}),$$
$$\widehat{\boldsymbol{x}}_t = F(\boldsymbol{y}_t, \boldsymbol{z}_t; \boldsymbol{\theta}),$$

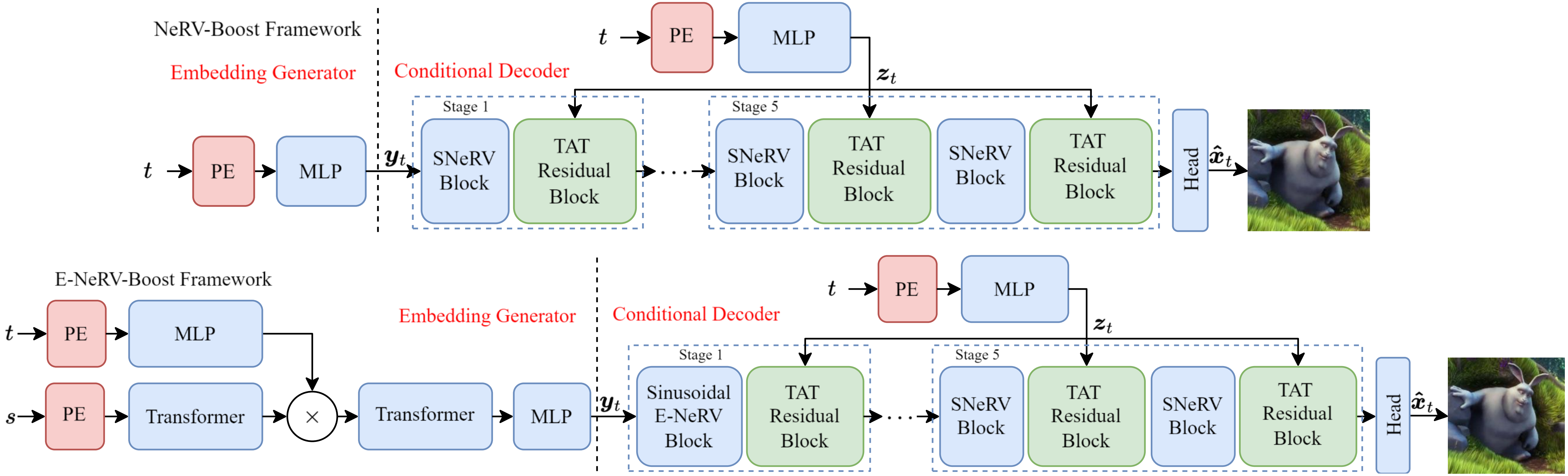
where $\mathrm{PE}(t) = \left(\sin(b^0 \pi t), \cos(b^0 \pi t), \ldots, \sin\left(b^{l-1} \pi t\right), \cos\left(b^{l-1} \pi t\right)\right)$



THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Method: Universal Boosting-NeRV Framework

Our boosting framework can be easily generalized to other representation models (e.g., NeRV, E-NeRV and so on) by selecting appropriate embedding generators.

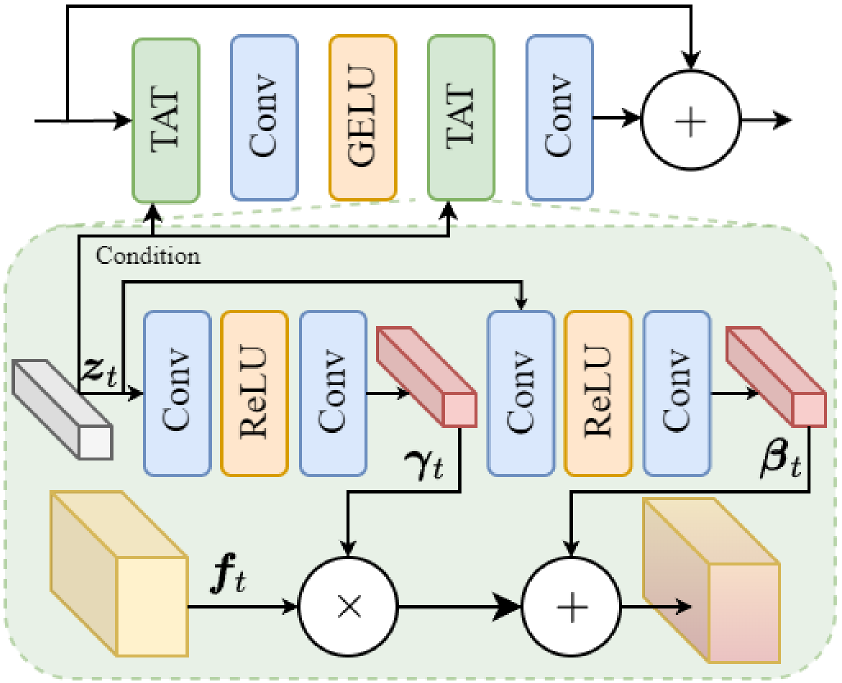


Method: Temporal-aware Affine Transform

The TAT layer takes the temporal embeddings $\boldsymbol{z_t}$ to produce channel-wise scaling and shifting parameters $\boldsymbol{\gamma_t}$ and $\boldsymbol{\beta_t}$.

$$\mathrm{TAT}(\boldsymbol{f}_t|\boldsymbol{\gamma}_t, \boldsymbol{\beta}_t) = \boldsymbol{\gamma}_t \boldsymbol{f}_t + \boldsymbol{\beta}_t,$$

By inserting the TAT residual block into existing video INRs, these aligned intermediate features can significantly enhance the models' overfitting ability



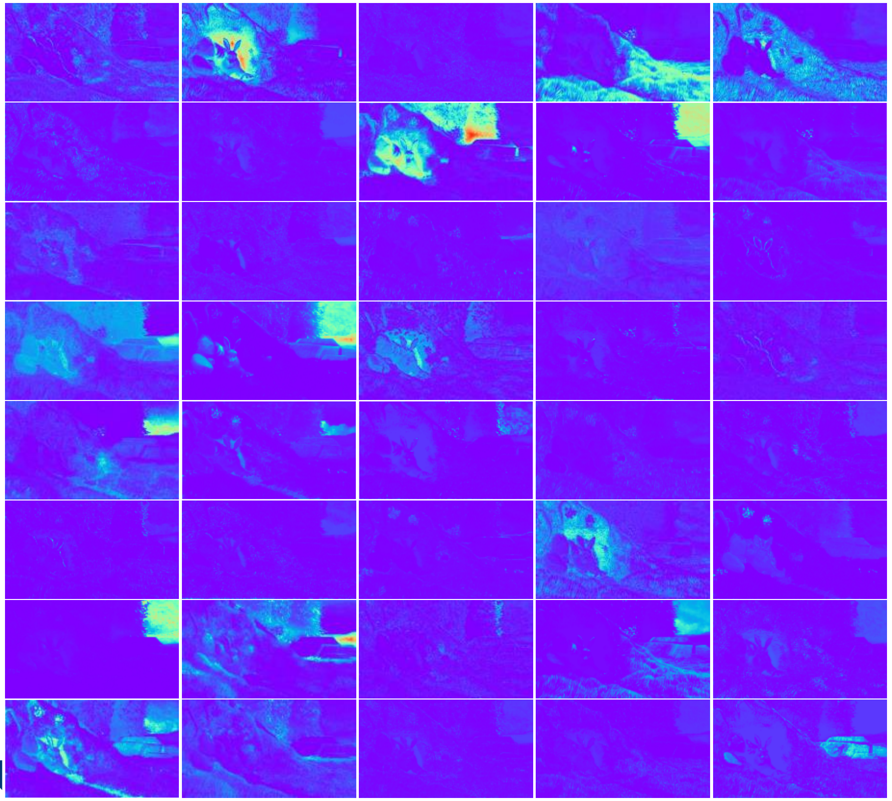
Identity-aware reconstruction!

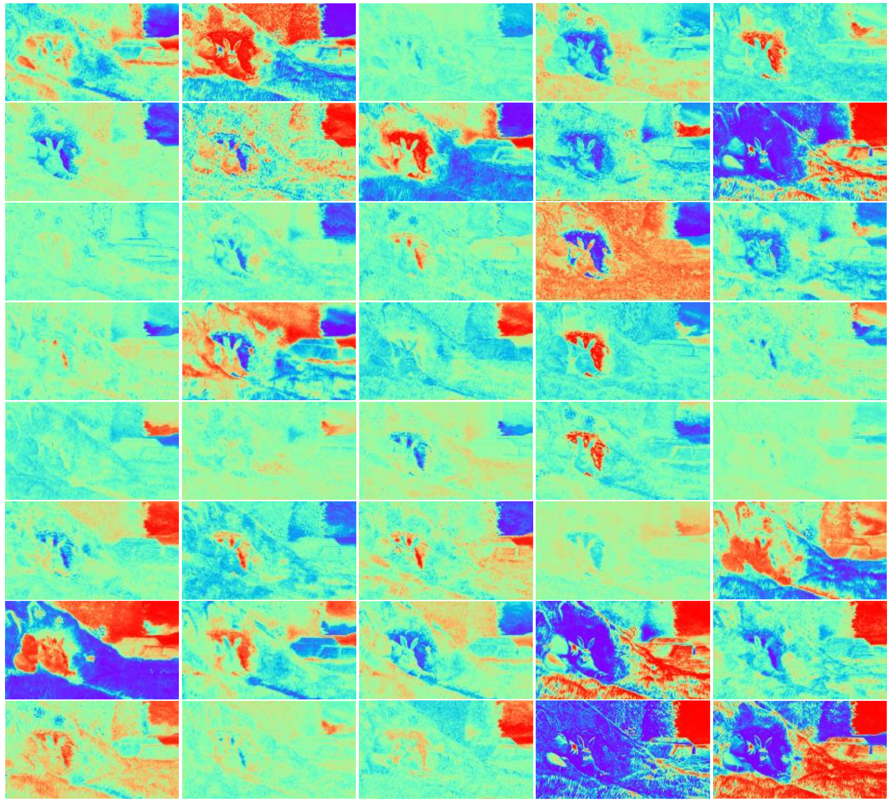
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Method: Sinusoidal NeRV-like Block

(Left) GELU: activate only a limited number of features.

(Right) SINE: activate diverse features & focus on different regions.





Diverse feature activation!

Method: Sinusoidal NeRV-like Block

Replace a single HNeRV block with a 5×5 kernel for two SNeRV blocks with a 3×3 kernel.

Table 9. Ablation studies for various upsampling blocks in the HNeRV-Boost framework on the Bunny video. GELU and SINE represent the activation function employed in different blocks. STD refers to the standard deviation of the model parameters' distribution, where a lower STD value signifies a more uniform distribution of model parameters.

Block	NeRV	E-NeRV	FFNeRV	HNeRV	SNeRV
GELU	39.61	39.26	39.33	40.77	**41.00**
SINE	40.35	39.99	40.06	40.93	**41.09**
STD	0.225	0.208	0.176	0.047	0.045

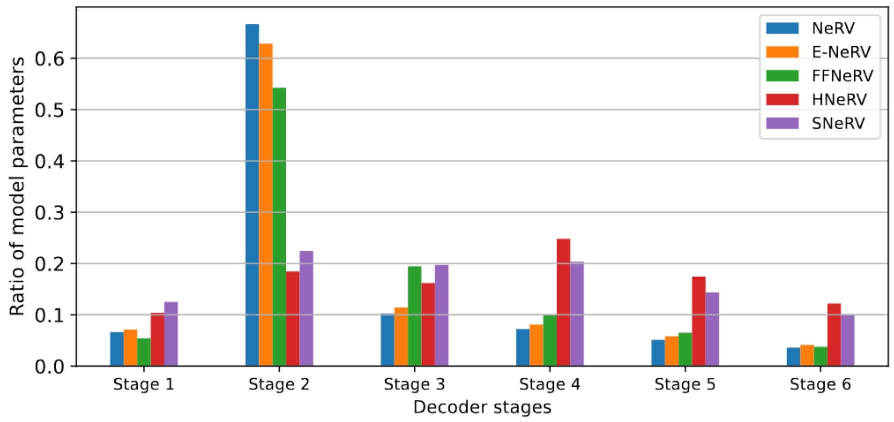


Figure 6. Distribution of model parameters across various decoder blocks in our HNeRV-Boost framework. See Table 9 for PSNR results under these five configurations.

Evenly-distributed Parameters!

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Method: High-frequency Supervision

Loss function: integrate a combination of the MS-SSIM and frequency domain losses into the L1 loss, ensuring a more comprehensive capture of high-frequency regions.

$$\mathcal{L}_d = \mathcal{L}_1(FFT(\boldsymbol{x}_t), FFT(\hat{\boldsymbol{x}}_t)) + \lambda\alpha\mathcal{L}_1(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_t)$$
$$+\lambda(1-\alpha)(1-\mathcal{L}_{MS-SSIM}(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_t))$$

Details-preserving reconstruction!

Method: Consistent Entropy Minimization

Symmetric Quantization:

$$Q(x) = \left\lfloor \frac{x}{\varsigma} \right\rceil, Q^{-1}(x) = x \times \varsigma,$$

Asymmetric Quantization:

$$Q(x) = \left\lfloor \frac{x-\eta}{\varsigma} \right\rceil, Q^{-1}(x) = x \times \varsigma + \eta,$$

Network-free Gaussian Entropy Model:

$$p(\hat{\boldsymbol{y}}_t) = \prod_i \left(\mathcal{N}(\mu_{\boldsymbol{y}_t}, \sigma^2_{\boldsymbol{y}_t}) * \mathcal{U}\left(-\frac{1}{2}, \frac{1}{2}\right)\right)(\hat{y}^i_t),$$

Optimization Objective:

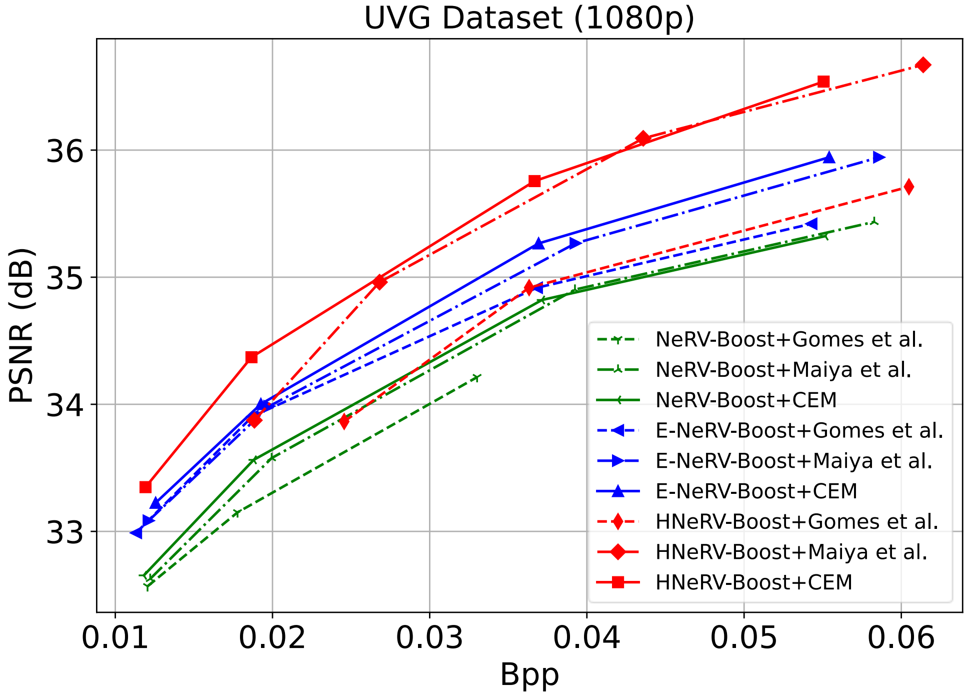
$$\mathcal{L} = \mathcal{L}_d + \kappa \mathcal{L}_r = \mathcal{L}_d + \kappa \mathrm{ReLU}(R - R_{target}),$$

$$R = \frac{\sum_{t=1}^{T} R(\hat{\boldsymbol{y}}_t) + R(\hat{\boldsymbol{\theta}}) + R(\hat{\boldsymbol{\psi}})}{T \times H \times W},$$

$$R_{target} = B_{avg} \frac{\sum_{t=1}^{T} Numel(\boldsymbol{y}_t) + Numbel(\boldsymbol{\theta}) + Numbel(\boldsymbol{\psi})}{T \times H \times W},$$

Table 1. Comparisons between different entropy minimization techniques in INR compression.

Method	Quantization		Entropy Model	
	Weight	Embedding	Training	Inference
Gomes *et al.*	Asymmetric	-	Neural network	CABAC
Maiya *et al.*	Symmetric	-	Neural network	Fixed frequency table
CEM (ours)	Symmetric	Asymmetric	Network-free Gaussian entropy model	



THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

[7] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, Nick Johnston. "Variational image compression with a scale hyperprior", ICLR 2018.

Comprehensive Evaluation: Video Regression

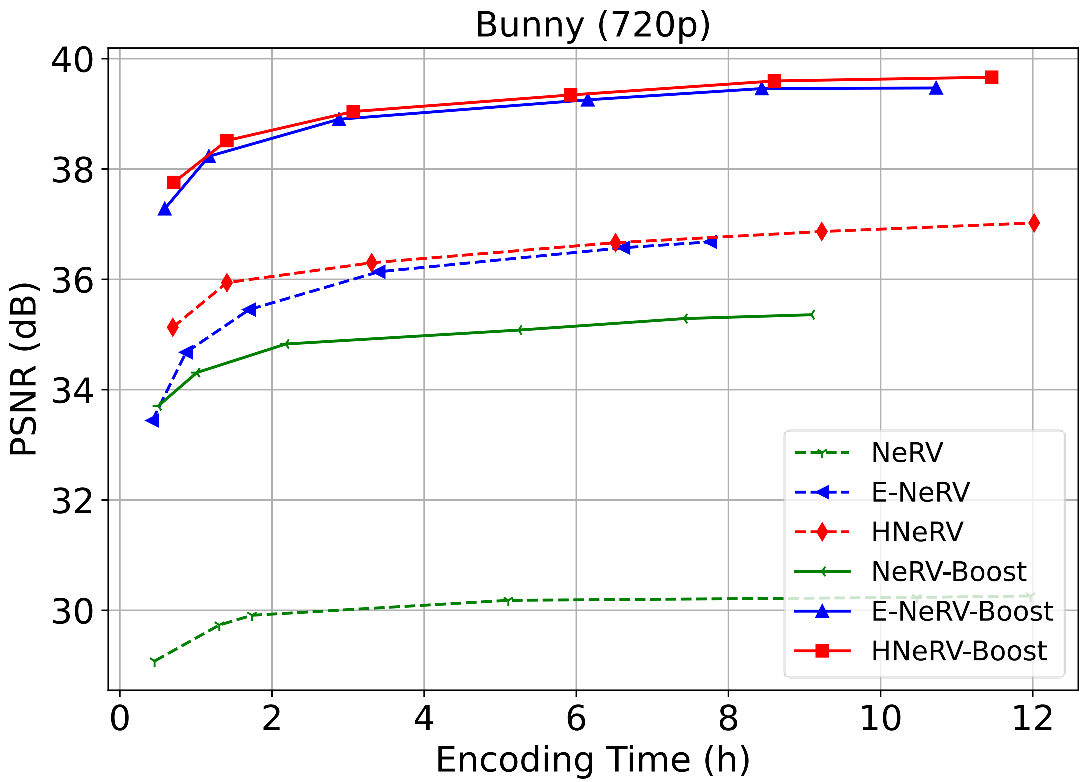


Table 4. PSNR on the Bosphorus video with different epochs.

Epoch	300	600	1200	1800	2400
NeRV	32.74	33.00	33.20	33.27	33.32
NeRV-Boost	**34.51**	**34.73**	**34.89**	**34.97**	**35.02**
E-NeRV	33.87	34.19	34.40	34.50	34.56
E-NeRV-Boost	**35.62**	**35.92**	**36.16**	**36.27**	**36.32**
HNeRV	33.62	34.15	34.35	34.41	34.46
HNeRV-Boost	**36.11**	**36.33**	**36.52**	**36.59**	**36.64**

Comprehensive Evaluation: Video Regression

Table 11. Video regression results on the UVG dataset in PSNR and MS-SSIM.

Model	Size	PSNR								MS-SSIM							
		Beauty	Bosph.	Honey.	Jockey	Ready.	Shake.	Yacht.	Avg.	Beauty	Bosph.	Honey.	Jockey	Ready.	Shake.	Yacht.	Avg.
NeRV	3.04M	33.14	32.74	37.18	30.99	23.97	33.06	26.72	31.11	0.8918	0.9358	0.9806	0.8989	0.8426	0.9347	0.8712	0.9079
NeRV-Boost	3.06M	**33.55**	**34.51**	**39.04**	**32.82**	**26.08**	**34.54**	**28.76**	**32.76**	**0.8967**	**0.9480**	**0.9840**	**0.9174**	**0.8768**	**0.9458**	**0.8931**	**0.9231**
E-NeRV	3.01M	33.29	33.87	38.88	28.73	23.98	34.45	27.38	31.51	0.8933	0.9444	0.9843	0.8708	0.8449	0.9468	0.8842	0.9098
E-NeRV-Boost	3.03M	**33.75**	**35.62**	**39.61**	**32.39**	**27.75**	**35.48**	**29.23**	**33.40**	**0.8987**	**0.9577**	**0.9854**	**0.9101**	**0.9057**	**0.9543**	**0.9015**	**0.9305**
HNeRV	3.05M	33.36	33.62	39.17	32.31	25.60	34.90	28.33	32.47	0.8907	0.9320	0.9843	0.8948	0.8490	0.9479	0.8642	0.9090
HNeRV-Boost	3.05M	**33.80**	**36.11**	**39.65**	**34.28**	**28.19**	**35.88**	**29.33**	**33.89**	**0.8996**	**0.9653**	**0.9854**	**0.9298**	**0.9139**	**0.958**	**0.9019**	**0.9363**
NeRV	5.07M	33.62	34.32	38.32	32.86	25.67	34.24	28.06	32.44	0.8994	0.9528	0.9832	0.9230	0.8854	0.9488	0.9015	0.9277
NeRV-Boost	5.00M	**33.89**	**35.86**	**39.31**	**34.16**	**27.78**	**35.33**	**30.00**	**33.76**	**0.9020**	**0.9608**	**0.9846**	**0.9332**	**0.9072**	**0.9564**	**0.9160**	**0.9372**
E-NeRV	5.09M	33.77	35.38	39.33	31.56	25.37	35.23	28.64	32.76	0.9002	0.9596	0.9851	0.9050	0.8804	0.9561	0.9098	0.9280
E-NeRV-Boost	5.01M	**34.02**	**36.79**	**39.71**	**33.90**	**29.29**	**36.20**	**30.24**	**34.31**	**0.9026**	**0.9669**	**0.9856**	**0.9287**	**0.9283**	**0.9626**	**0.9181**	**0.9418**
HNeRV	5.06M	33.84	34.49	39.56	33.64	27.24	35.73	29.29	33.40	0.8987	0.9430	0.9853	0.9114	0.8848	0.9588	0.8857	0.9240
HNeRV-Boost	5.01M	**34.14**	**37.87**	**39.74**	**35.84**	**30.36**	**36.71**	**30.77**	**35.06**	**0.9045**	**0.9764**	**0.9857**	**0.9467**	**0.9413**	**0.9675**	**0.9249**	**0.9496**
NeRV	10.10M	34.10	36.52	39.35	35.37	28.10	35.82	30.11	34.20	**0.9088**	0.9701	0.9852	0.9493	0.9302	0.9662	0.9354	0.9493
NeRV-Boost	10.08M	**34.17**	**37.77**	**39.65**	**36.23**	**30.25**	**36.81**	**32.06**	**35.28**	0.9074	**0.9749**	**0.9855**	**0.9525**	**0.9419**	**0.9703**	**0.9429**	**0.9536**
E-NeRV	10.16M	34.18	37.31	39.70	34.62	28.27	36.50	30.36	34.42	0.9065	0.9733	0.9858	0.9396	0.9297	0.9689	0.9361	0.9486
E-NeRV-Boost	10.04M	**34.28**	**38.39**	**39.82**	**35.88**	**31.42**	**37.34**	**31.94**	**35.58**	**0.9065**	**0.9767**	**0.9859**	**0.9481**	**0.9515**	**0.9730**	**0.9389**	**0.9544**
HNeRV	10.07M	34.22	37.27	39.73	34.59	29.59	36.82	30.70	34.70	0.9053	0.9695	0.9857	0.9215	0.9255	0.9696	0.9134	0.9415
HNeRV-Boost	10.03M	**34.42**	**39.75**	**39.83**	**37.57**	**33.12**	**37.85**	**32.90**	**36.49**	**0.9096**	**0.984**	**0.9859**	**0.9617**	**0.9647**	**0.9768**	**0.9475**	**0.9615**
NeRV	15.09M	34.36	37.66	39.59	36.55	29.81	36.86	31.43	35.18	**0.9170**	0.9766	**0.9857**	0.9588	**0.9509**	0.9741	0.9508	0.9591
NeRV-Boost	15.04M	**34.47**	**38.87**	**39.67**	**37.35**	**30.87**	**37.37**	**33.00**	**35.94**	0.9157	**0.9803**	0.9855	**0.9622**	0.9481	**0.9742**	**0.9527**	**0.9598**
E-NeRV	15.02M	34.34	38.41	39.78	35.98	29.90	37.32	31.52	35.32	**0.9111**	0.9790	0.9860	0.9528	0.9492	0.9754	0.9491	0.9575
E-NeRV-Boost	15.06M	**34.40**	**39.31**	**39.85**	**36.90**	**32.46**	**37.99**	**33.00**	**36.27**	0.9089	**0.9819**	**0.9860**	**0.9574**	**0.9598**	**0.9776**	**0.9496**	**0.9602**
HNeRV	15.02M	34.37	38.40	39.81	35.76	31.02	37.00	31.82	35.45	0.9079	0.9766	0.9859	0.9370	0.9435	0.9705	0.9295	0.9501
HNeRV-Boost	15.04M	**34.65**	**40.72**	**39.88**	**38.41**	**34.72**	**38.47**	**34.16**	**37.29**	**0.9176**	**0.9870**	**0.9861**	**0.9678**	**0.9739**	**0.9803**	**0.9578**	**0.9672**

Comprehensive Evaluation: Video Compression

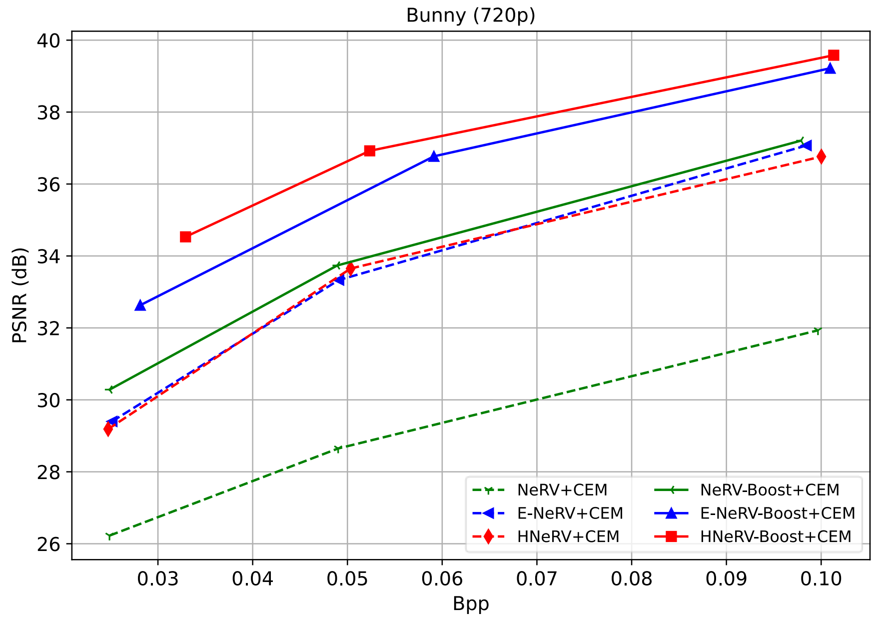
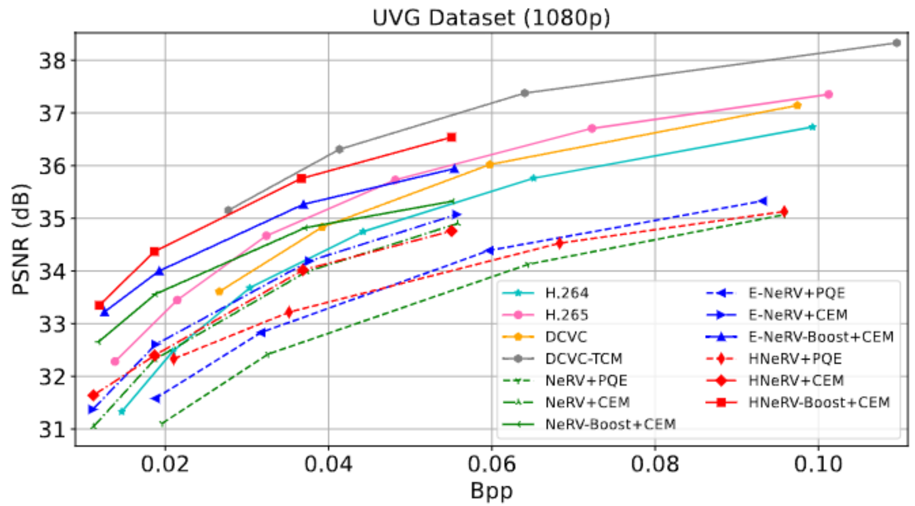


Table 5. Complexity comparison at resolution 1920×1080. The decoding latency is evaluated by an NVIDIA V100 GPU.

Method	Params ↓	Decoding time ↓	FPS ↑
DCVC	35.2M	35590ms	0.028
DCVC-TCM	40.9M	470ms	2.12
NeRV	3.04M	7ms	135.64
NeRV-Boost	3.06M	23ms	43.54
E-NeRV	3.01M	18ms	54.75
E-NeRV-Boost	3.03M	53ms	18.74
HNeRV	3.05M	41ms	24.22
HNeRV-Boost	3.06M	76ms	13.15



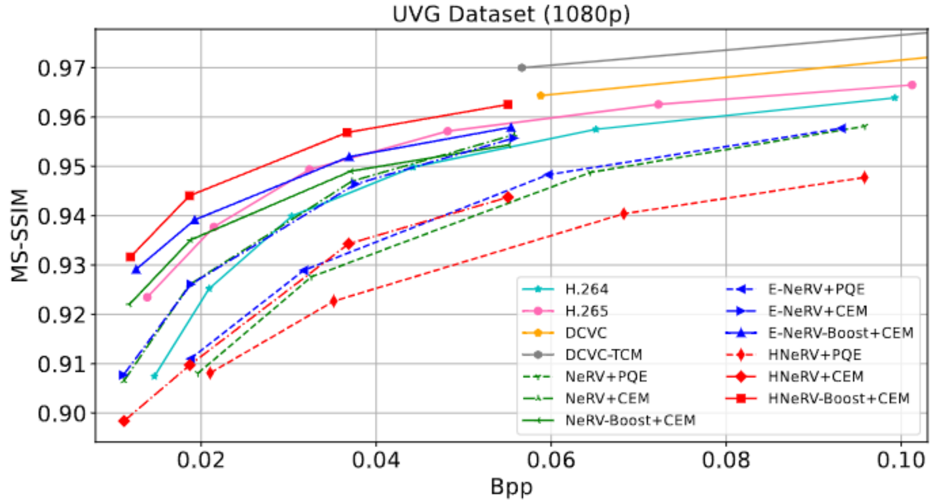


Figure 5. Rate-distortion curves of our boosted approaches and different baselines on the UVG dataset in PSNR and MS-SSIM. PQE denotes the three-step compression pipeline of NeRV.

Comprehensive Evaluation: Video Inpainting

Table 6. Video inpainting results on the DAVIS validation dataset in PSNR. Mask-S and Mask-C refers to disperse and central mask scenarios, respectively.

Video	Mask-S						Mask-C					
	NeRV	NeRV-Boost	E-NeRV	E-NeRV-Boost	HNeRV	HNeRV-Boost	NeRV	NeRV-Boost	E-NeRV	E-NeRV-Boost	HNeRV	HNeRV-Boost
Blackswan	27.06	**30.46**	29.53	**31.34**	30.20	**34.10**	24.11	**26.89**	26.38	**27.88**	26.45	**29.18**
Bmx-trees	26.77	**30.16**	27.75	**30.86**	29.05	**32.99**	22.43	**25.14**	23.79	**26.66**	22.28	**22.28**
Breakdance	25.48	**28.46**	26.97	**30.57**	26.34	**33.10**	20.16	**22.28**	22.15	**22.15**	20.23	**20.24**
Camel	23.70	**26.09**	25.70	**27.56**	26.13	**31.08**	21.21	**23.16**	22.62	**23.55**	17.74	**19.81**
Car-roundabout	23.92	**28.25**	26.32	**29.43**	28.64	**31.90**	21.24	**23.53**	22.73	**24.51**	21.71	**22.36**
Car-shadow	26.58	**32.40**	30.63	**33.00**	31.01	**35.85**	23.07	**24.13**	23.21	**24.10**	21.05	**23.65**
Cows	22.17	**24.77**	23.92	**26.41**	24.68	**28.30**	20.48	**22.39**	21.88	**23.13**	21.82	**24.14**
Dance-twirl	25.29	**28.49**	27.42	**29.38**	28.74	**30.79**	21.17	**23.14**	22.40	**23.34**	21.06	**21.77**
Dog	29.29	**31.97**	31.72	**32.79**	28.80	**33.87**	25.37	**27.02**	27.07	**28.25**	24.16	**24.66**
Drift-chicane	34.09	**39.94**	39.26	**41.60**	38.52	**43.32**	27.52	**28.01**	29.81	**31.52**	23.40	**27.44**
Drift-straight	26.78	**32.26**	29.53	**33.19**	30.81	**36.16**	22.76	**26.00**	24.69	**27.12**	18.88	**21.49**
Goat	24.04	**26.30**	25.34	**27.21**	26.91	**30.59**	22.03	**23.90**	23.43	**24.56**	23.06	**25.10**
Horsejump-high	25.74	**30.39**	29.27	**31.26**	29.31	**30.86**	21.54	**23.46**	23.06	**23.93**	20.72	**23.16**
Kite-surf	29.34	**34.18**	32.87	**35.16**	33.49	**37.08**	23.92	**27.22**	26.71	**28.87**	24.73	**27.49**
Libby	29.81	**34.24**	31.39	**34.95**	28.66	**37.35**	25.71	**28.14**	26.91	**28.95**	23.39	**26.96**
Motocross-jump	29.82	**37.36**	34.15	**36.92**	28.27	**36.42**	26.19	**29.65**	28.75	**29.30**	22.36	**26.25**
Paragliding-launch	29.03	**31.40**	30.62	**32.28**	30.99	**33.64**	25.95	**26.97**	26.65	**27.41**	26.00	**28.07**
Parkour	24.74	**27.19**	25.62	**27.54**	26.34	**28.79**	22.32	**24.48**	22.99	**24.43**	19.06	**20.55**
Scooter-black	23.35	**27.75**	26.46	**29.07**	28.41	**30.42**	19.24	**21.77**	20.99	**22.14**	18.94	**19.86**
Soapbox	27.20	**30.56**	28.83	**31.44**	30.30	**32.95**	22.29	**25.00**	23.82	**25.51**	17.98	**19.20**
Average	26.71	**30.63**	29.17	**31.60**	29.28	**33.48**	22.94	**25.11**	24.50	**25.87**	21.75	**23.68**

Comprehensive Evaluation: Video Interpolation

Table 7. Video interpolation results on the UVG dataset in PSNR.

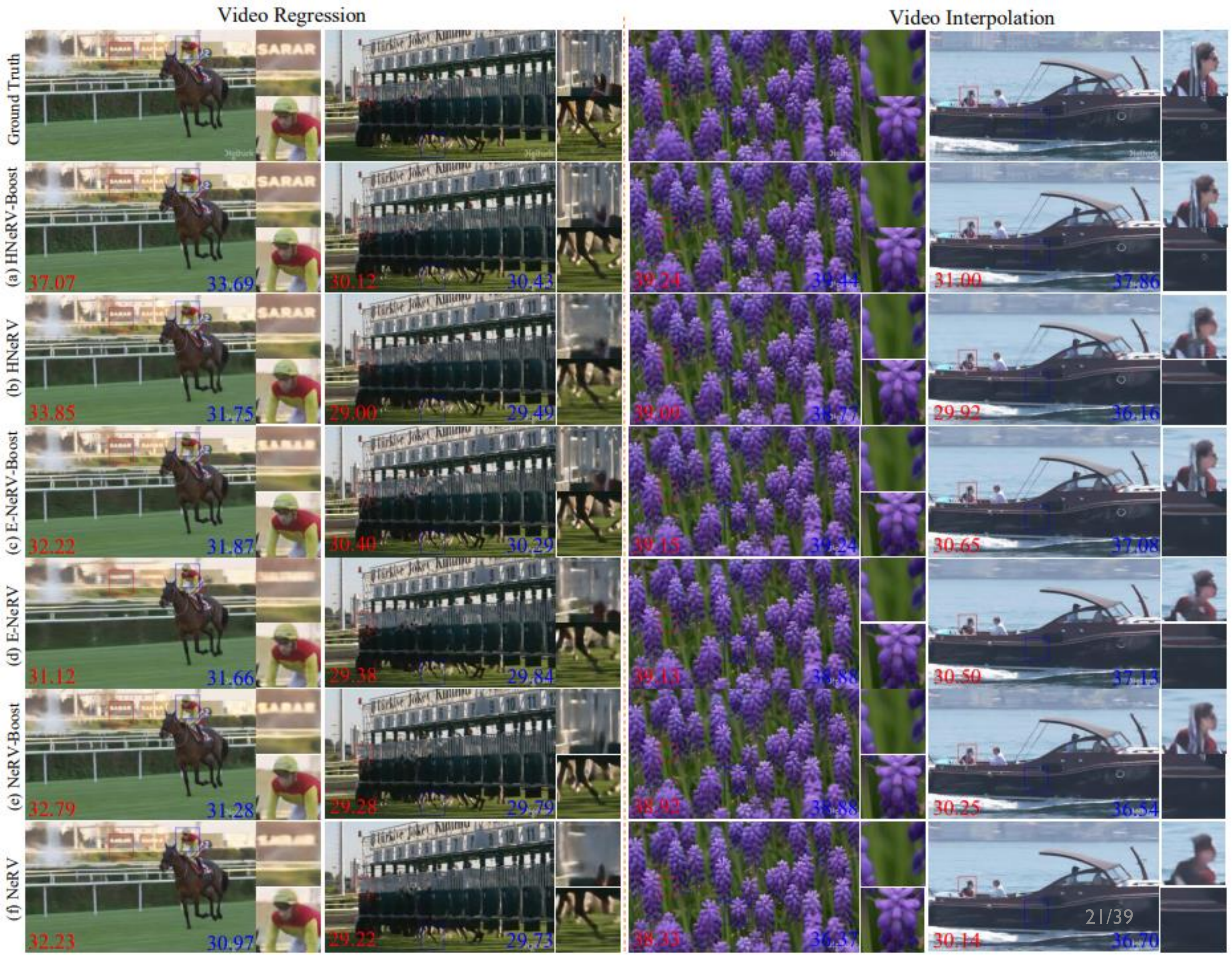
Video	Beauty	Bosph.	Honey.	Jockey	Ready.	Shake.	Yacht.	Avg.
NeRV	**31.26**	32.21	36.84	**22.24**	**20.05**	32.09	26.09	28.68
NeRV-Boost	31.06	**34.28**	**38.83**	21.74	19.88	**32.58**	**27.07**	**29.35**
E-NeRV	31.25	33.36	38.62	**22.35**	20.08	**32.82**	26.74	29.32
E-NeRV-Boost	**31.35**	**35.01**	**39.24**	21.96	**20.45**	32.75	**27.79**	**29.79**
HNeRV	31.42	34.00	39.07	23.02	20.71	32.58	26.74	29.65
HNeRV-Boost	**31.61**	**36.16**	**39.38**	**23.14**	**21.61**	**32.94**	**28.01**	**30.41**

Table 12. Video interpolation results on the UVG dataset in MS-SSIM.

Model	Beauty	Bosph.	Honey.	Jockey	Ready.	Shake.	Yacht.	Avg.
NeRV	**0.8696**	0.9297	0.9797	**0.7542**	**0.7070**	0.9238	0.8578	0.8603
NeRV-Boost	0.8673	**0.9461**	**0.9835**	0.7357	0.6970	**0.9250**	**0.8708**	**0.8608**
E-NeRV	0.8702	0.9383	0.9838	**0.7536**	0.7029	**0.9288**	0.8720	0.8642
E-NeRV-Boost	**0.8719**	**0.9525**	**0.9846**	0.7476	**0.7233**	0.9272	**0.8857**	**0.8704**
HNeRV	0.8702	0.9379	0.9841	0.7677	0.7056	0.9263	0.8287	0.8601
HNeRV-Boost	**0.8754**	**0.9664**	**0.9849**	**0.7836**	**0.7527**	**0.9284**	**0.8897**	**0.8830**

Visualizations

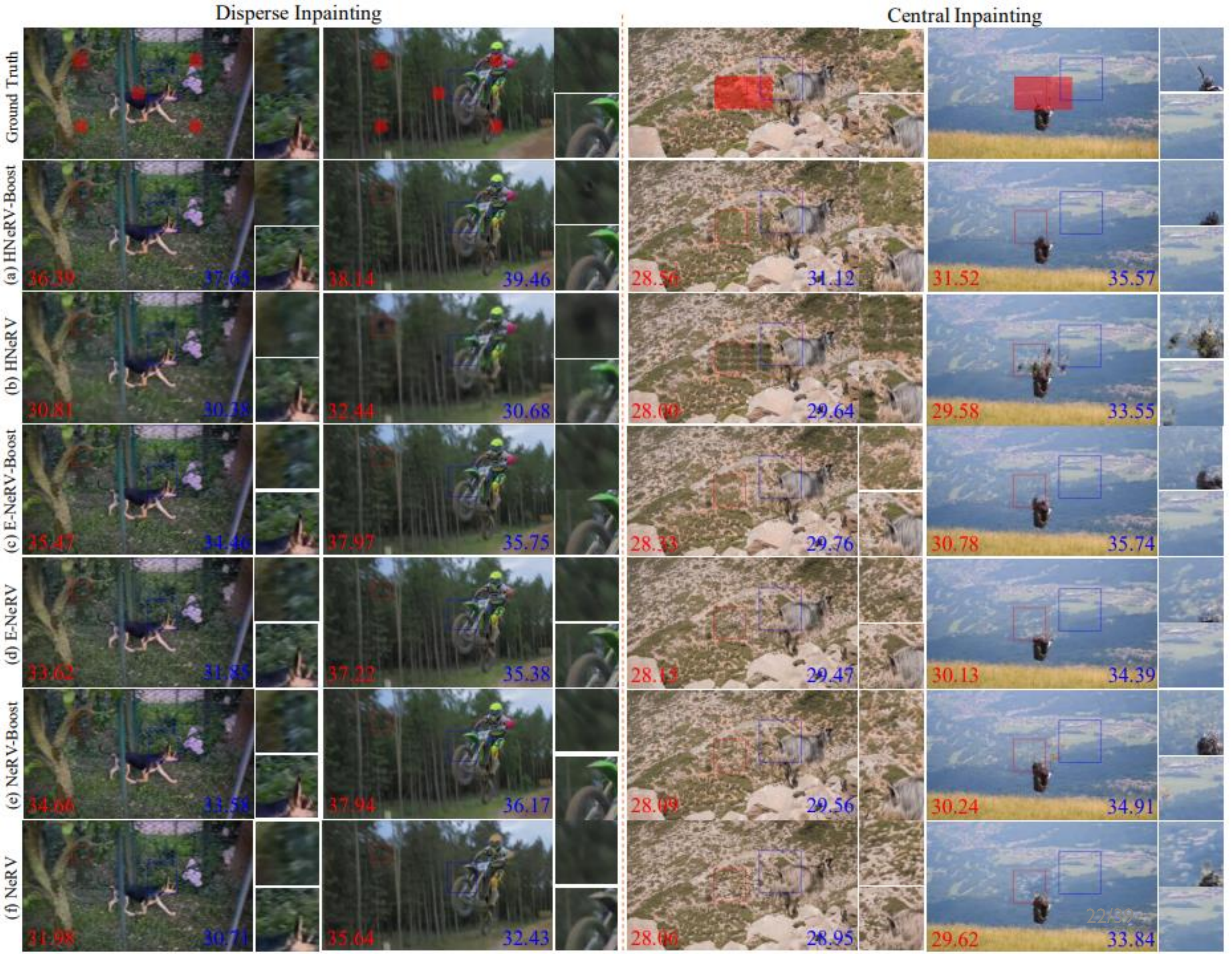
Qualitative results of video regression and interpolation.



THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Visualizations

Qualitative results of video inpainting.



THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Ablation Study

Table 8. Ablation studies for different boosting components on the Bunny video over 300 epochs, with results presented in PSNR.

Variant	NeRV-Boost	E-NeRV-Boost	HNeRV-Boost
Ours	**37.25**	**40.07**	**41.09**
(V1) w/o TAT	34.63	35.75	39.12
(V2) w/ AdaIN	35.59	39.51	38.03
(V3) w/ SAF	34.28	39.62	40.93
(V4) w/ GELU	34.85	38.34	41.00
(V5) w/ L2	35.32	38.55	40.19
(V6) w/ L1+SSIM	36.28	39.34	41.00
(V7) w/ L1	34.75	37.76	40.37
(V8) w/ L1+MS-SSIM	36.12	38.66	40.49
(V9) w/ L1+freq.	37.12	39.60	41.08
(V10) w/ L1+SSIM+freq.	36.99	40.05	41.01

Table 15. Ablation study of different boosting components on the Buuny video with 3M model size and 300 training epochs.

Model	TAT	SNeRV	Improved loss	PSNR
NeRV				31.84
	✓			33.50
	✓	✓		36.28
	✓	✓	✓	**37.25**
E-NeRV				37.32
	✓			38.01
	✓	✓		39.34
	✓	✓	✓	**40.07**
HNeRV				38.15
	✓			39.90
	✓	✓		40.19
	✓	✓	✓	**41.09**

Summary

- Develop a universal framework to boost existing NeRV models, setting a new benchmark in the field of implicit video representation.

- These advancements are primarily due to the integration of several novel developments, including the temporal-aware affine transform, sinusoidal NeRV-like block design, improved reconstruction loss, and consistent entropy minimization.

- Source Code: https://github.com/Xinjie-Q/Boosting-NeRV

Thank you!

香港科技大學
THE HONG KO G
UNIVERSITY OF SCIENCE
AND TECHNOLOGY