Learning-based Video Compression

From TV to the Metaverse

Prof. David Bull

Visual Information Lab. and Director MyWorld, University of Bristol

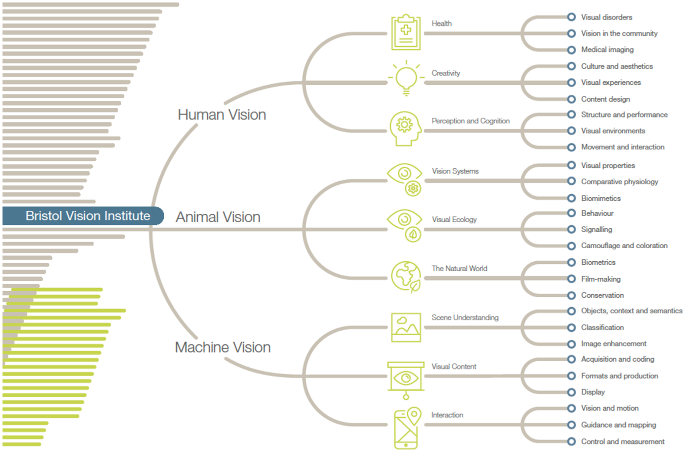
9 Dec 2024 @ Tokyo, Japan

About Me

- **Professor** in Signal Processing, UoB
- **Founder Director**, Bristol Vision Institute
- **Director**, MyWorld, UK Strength in Places Fund
- **Author**, Bull and Zhang, *Intelligent Image and Video Compression*, Academic Press, 2021

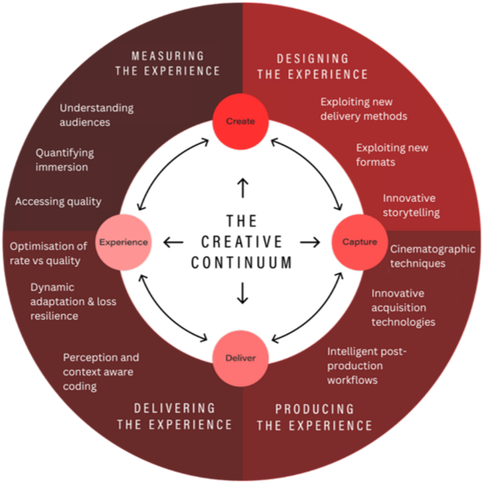
Bristol Vision Institute

- Formed in **2008**.
- Hosting some **160** researchers.
- An intellectual landscape and practical facilities for **vision research**.
- Facilitates **engineers** and **scientists** working together with experts in **medicine** and **creative arts**.
- **One of the largest inter-disciplinary** groups in Europe.
- Successful - attracting **research income**, stimulating new relationships and creating **commercial impact**.



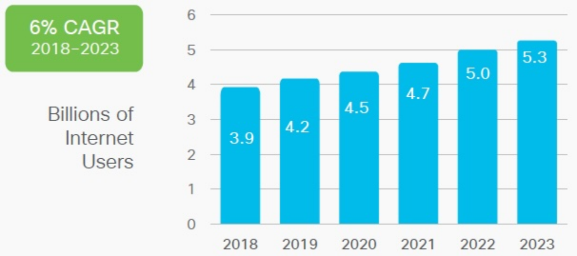
MyWorld

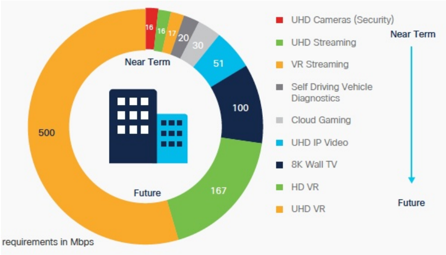
- A **£30m** investment under the UKRI Strength in Places Fund. Exploiting the production, technology and research strengths of the **West of England's** creative sector.
- **25** new major **international partnerships**.
- Additional funding leveraged ~**£29M**.
- **368** businesses supported to date.
- **298** jobs created.
- **112,000** members of public engaged.
- **2036** individual learners.
- **22** awards, prizes and prestigious lectures.
- **129** academic outputs.



The Challenges of Video Compression

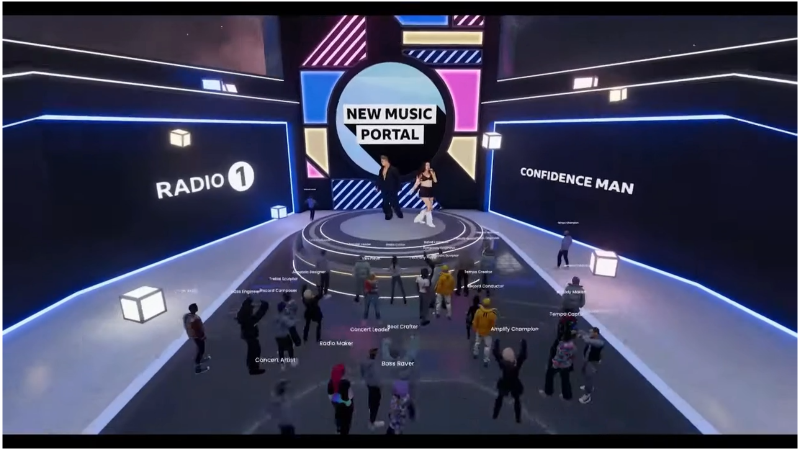
- Huge amounts of video content consumed via steaming and social media: e.g. **NETFLIX** and **TikTok**.
- Significantly increased demand for more **immersive services**, e.g. **UHD/HFR/HDR, XR and 360**°.
- Consistent growth in the number of the **global Internet users** - 5.3bn in 2023.





Source of figures: Cisco Annual Internet Report 2018-2023.

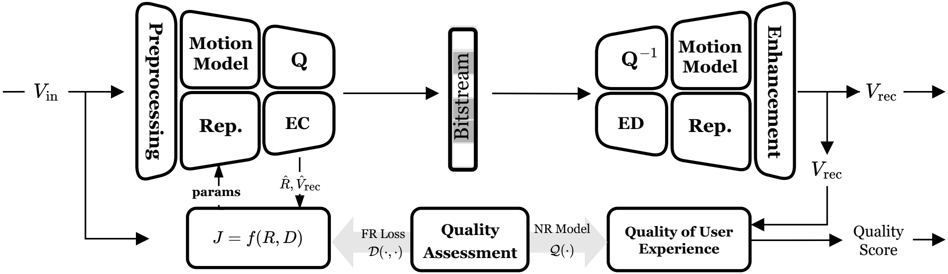
Example: Real-time Volumetric Video Delivery



[VIDEO] Live volumetric video delivery into the metaverse (`https://condense.live`).

A Video Compression Framework

- **Motion model**: motion estimation/compensation, advanced motion models, optical flows.
- **Representation**: transforms, feature extraction.
- **Quantisation and entropy coding**: data compression for residual, latent or models.
- **Enhancement**: pre- and post-processing, super resolution.
- **Quality assessment**: for rate-distortion optimisation (encoder) or QoE prediction.



Overview

Video Compression - pre AI

AI-based Video Compression

Reducing Complexity

Motion Models

Representation Models

Conclusion

Outline

Video Compression - pre AI

AI-based Video Compression

Reducing Complexity

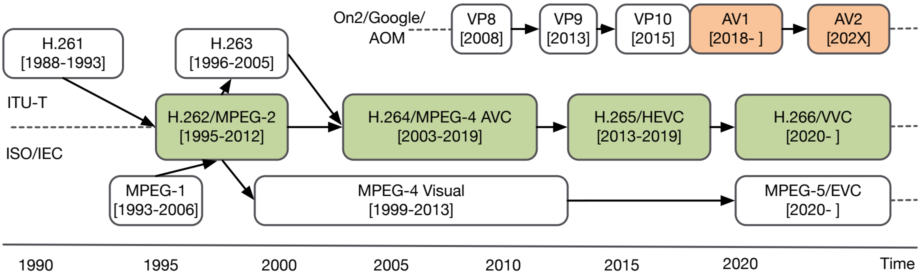
Motion Models

Representation Models

Conclusion

Video Coding Standards

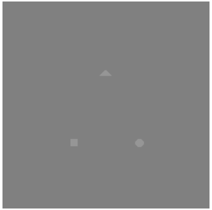
- VVC **VTM** achieves an average 29% bit rate saving against AOM **AV1**.
- The latest MPEG JVET test model **ECM** outperforms VTM by more than 25% in BD-rate saving.
- The new AOM codec **AVM** offers a 20%+ coding gain over AV1 libaom.



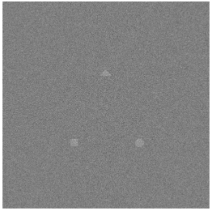
[Nguyen and Marpe, 2021] "Compression efficiency analysis of AV1, VVC, and HEVC for random access applications", APSIPA Transactions on Signal and Information Processing.

[Seregin *et al.*, 2024] "JVET AHG report: ECM software development (AHG6)", JVET-AI0006.

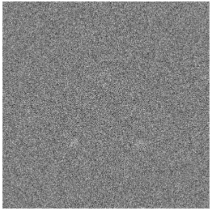
Textures and Video Coding



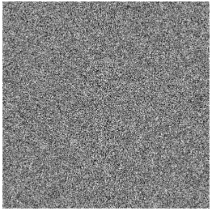
Three cues in
a grey background



With Gaussian noise
$(\mu = 0, \sigma = 0.001)$



With Gaussian noise
$(\mu = 0, \sigma = 0.0.01)$

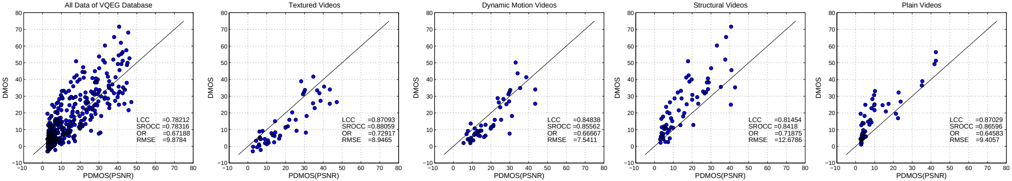


With Gaussian noise
$(\mu = 0, \sigma = 0.03)$

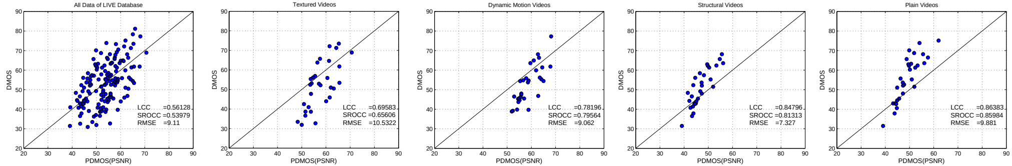
Quantisation Parameter (QP)	22	27	32	37	42
Static textures (bpp)	0.0278	0.0111	0.0051	0.0025	0.0012
Mixed textures (bpp)	0.2301	0.0684	0.0287	0.0133	0.0066
Dynamic textures (bpp)	0.3463	0.1904	0.0969	0.0473	0.0235

HEVC HM 16.4; Main Profile; Random access mode; BVI-Texture; 300 frames encoded.

Correlation between MSE/PSNR and Subjective Scores

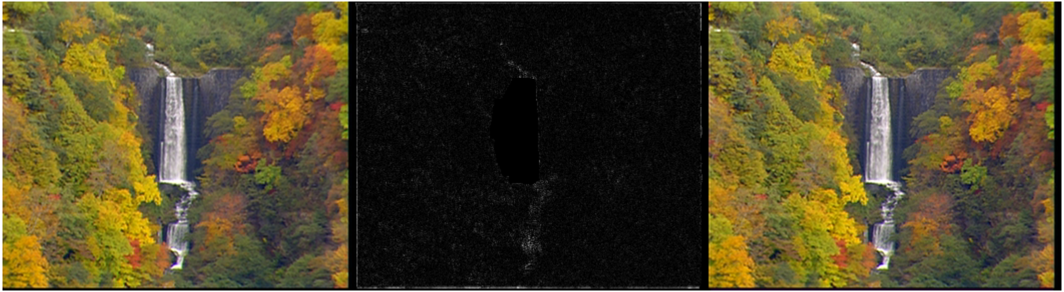


(a) All data:VQEG (b) Textured content (c) Dynamic content (d) Structural content (e) Luminance plain videos



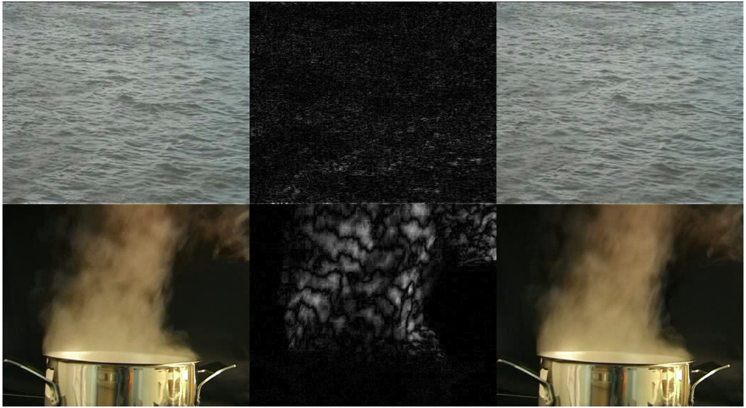
(f) All data: LIVE (g) Textured content (h) Dynamic content (i) Structural content (j) Luminance plain videos

Textures and Video Coding - Static Textures



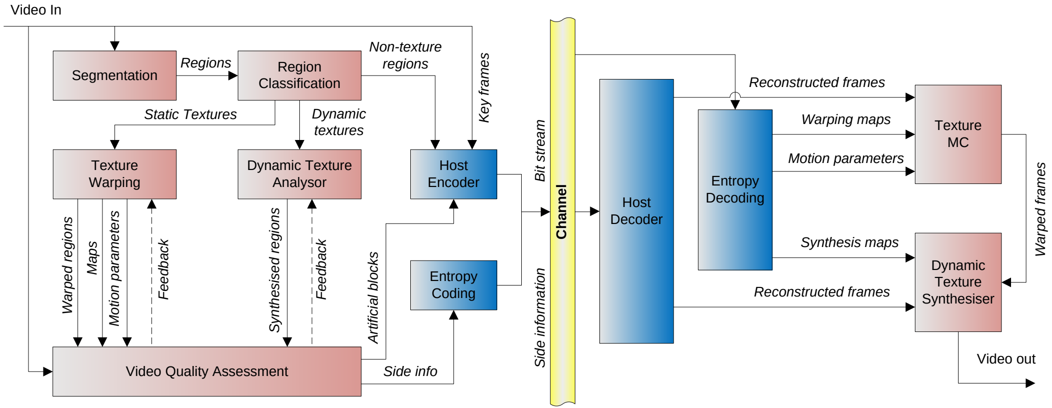
[VIDEO] **Left**: original texture. **Right**: warped texture. **Middle**: Absolute difference between left and right.

Textures and Video Coding - Dynamic Textures



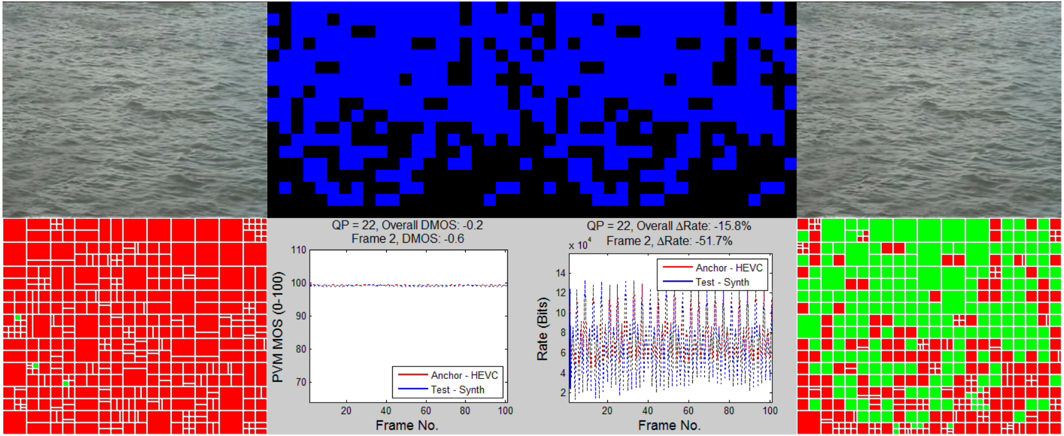
[VIDEO] **Left**: original texture. **Right**: synthesised texture. **Middle**: Absolute difference.

An Analysis-Synthesis Video Compression Framework



[Zhang and Bull, 2011] "A parametric framework for video compression using region-based texture models", IEEE Journal of Selected Topics in Signal Processing.

Compression Results based on HEVC



[VIDEO] **Left**: HEVC; **Right**: HEVC+Synthesis; **Middle**: Synthesis maps and RD stats.

Outline

Video Compression - pre AI

AI-based Video Compression

Reducing Complexity

Motion Models

Representation Models

Conclusion

Deep Video Compression: Overview

Background

- **Deep neural networks** now offer tractable solutions to many image processing problems.
- They are being increasingly applied in **image/video compression**, demonstrating significant coding gains.
- But often at the expense of increased complexity or latency.

AI-based video compression

- **Training databases**.
- **Deep video coding tools** for standard codec enhancement:
  e.g., post processing, in-loop filtering and resolution adaptation.
- **End-to-end learned video codecs**: e.g., DVC, DCVC codecs.
- **Perceptual quality assessment.**

Deep Video Compression: Training Databases

Motivation

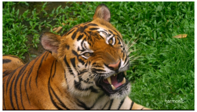
- DVC demands **volumes of training materiel** much greater than other machine learning methods.
- They must include **diverse content** covering different formats and video texture types.
- Most learning-based coding methods are currently trained on databases designed for image/video processing or computer vision applications.
- These training databases cannot ensure **network generalisation** or **optimum performance** for DVC.

Popular training databases for DVC

- **DIV2K** [Agustsson *et al.*, 2019]: contains 1000 RGB images and was developed for super-resolution.
- **CD** [Liu *et al.*, 2017]: collects 29 video sequences from LIVE VQA, MCL-V and TUM 1080p.
- **REDS** [Nah *et al.*, 2019a]: contains 300 video clips, and was developed for super-resolution.
- **Video Set** [Wang *et al.*, 2017]: includes 880 source videos, and was developed for quality assessment.
- **HIF** [Li *et al.*, 2019]: contains 182 video sequences, and was developed for deep video coding.

BVI-DVC: A Training Database for Deep Video Compression

- **BVI-DVC** contains 800 10bit video sequences at various spatial resolutions from 270p to 2160p.
- It covers **various video texture types**, including static textures and dynamic textures.


(a) Animal

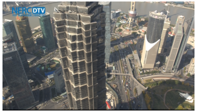

(b) Wood


(c) Leaves


(d) Mountain


(e) Myanmar


(f) Venice

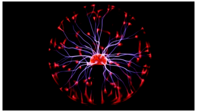

(g) Tall Buildings


(h) Traffic


(i) Market


(j) Ferris Wheel


(k) Cross Walk


(l) Plasma


(m) Firewood


(n) Smoke

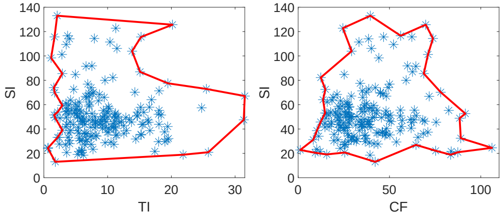

(o) Water

[Ma *et al.*, 2021] D. Ma et al., BVI-DVC: A Training Database for Deep Video Compression, IEEE Trans. in Multimedia, 2021.

Feature Coverage and Distribution

Training Databases	Image or Video?	Seq Number	Max Resolution	Bit Depth	Texture Diversity?
DIV2K [Agustsson and Timofte, 2017]	Image	1000	1152p	8	No
CD [Liu *et al.*, 2017]	Video	29	1080p	8	No
VideoSet [Wang *et al.*, 2017]	Video	880	1080p	8	No
REDS [Nah *et al.*, 2019b]	Video	300	720p	8	No
HIF [Li *et al.*, 2019]	Video	182	1080p	8	No
BVI-DVC	Video	800	2160p	10	Yes

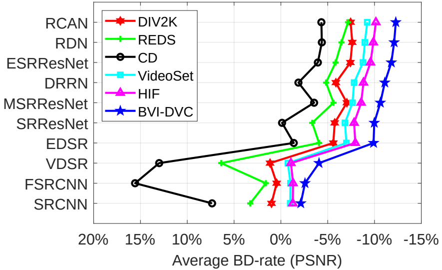
- Features [Winkler, 2012]:
- **SI** - spatial information.
- **TI** - temporal information.
- **CF** - colourfulness.

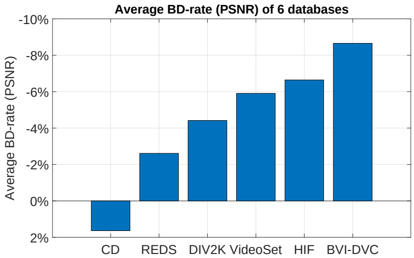


[Winkler, 2012] "Analysis of public image and video databases for quality assessment", IEEE Journal of Selected Topics in Signal Processing.

BVI-DVC vs Existing Training Databases for DVC

- **BVI-DVC** has been compared to five databases for DVC: **DIV2K**, **REDS**, **CD**, **Video Set** and **HIF**.
- The evaluation was conducted for four CNN-based coding tools based on **HEVC HM 16.20** and **JVET CTC**.
- **Ten popular network architectures** were used for evaluation.
- The coding gains were calculated **against the original HEVC HM**.

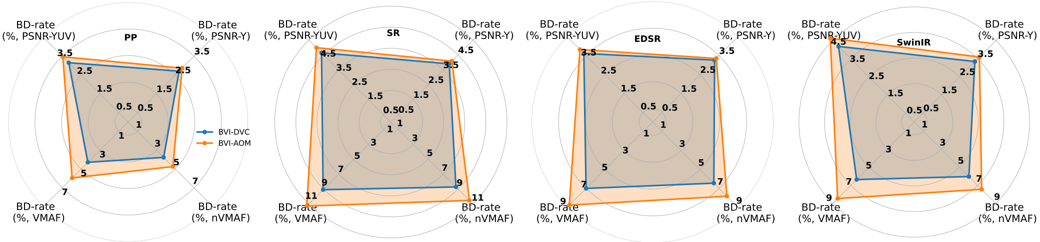




- BVI-DVC is **used in MPEG JVET** for developing VVC neural network based tools .

BVI-AOM

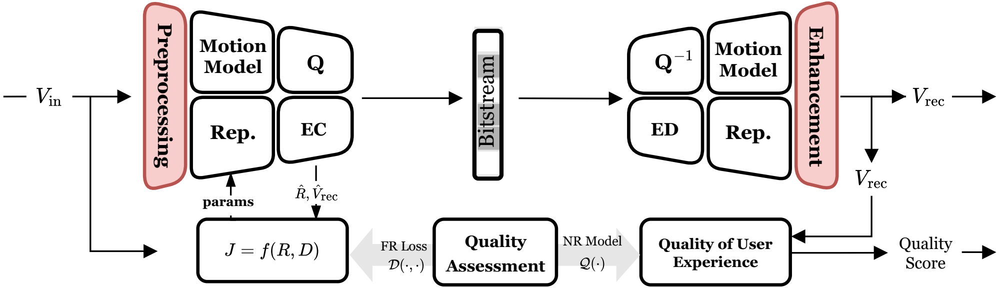
- BVI-AOM extends BVI-DVC with additional content, e.g. **dark** or **high-contrast** scenes.
- BVI-AOM offers **improved performance** (up to 2.98p.p.), with **more flexible licensing terms**.
- A collaboration with **Netflix (US)**, the database is available for **public downloading**.
- **Experimental setup**: two coding tools, two networks, four quality metrics and AOM CTCs.



[Nawała *et al.*, 2024] “BVI-AOM: A New Training Dataset for Deep Video Compression Optimization”, IEEE VCIP.

Conventional Coding Tools Enhancement

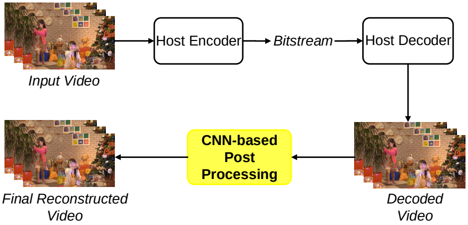
- Deep learning techniques have been applied to the improve coding efficiencey of various **existing coding tools**.
- Offering better performance when integrated into the **enhancement** modules.



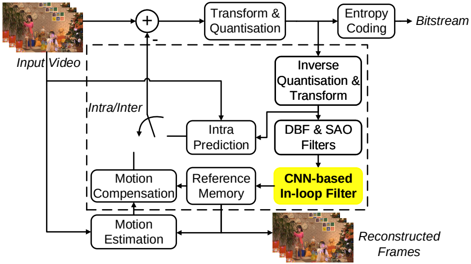
Enhancement of Coding Tools

- **Post-processing (PP)** and **in-loop filtering (ILF)** provide more consistent coding gains compared to other coding modules.

Post-processing (PP)

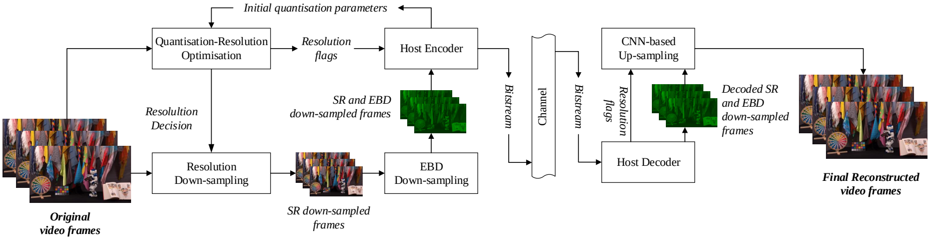


In-loop filtering (ILF)



ViSTRA: A Coding Framework based on Deep Learning

- ViSTRA trades off the relationship between **resolutions and quantisation** within the coding loop.
- Adaptation for spatial resolution (**SRA**), frame rate (for HFR only) and effective bit depth (**EBDA**).
- Resolution up-sampling is achieved through **CNN-based super resolution** (MSRResNet).
- **Machine learning inspired QRO**: spatial resolution adaptation based on quantisation and video content.



[Afonso *et al.*, 2018] "Video Compression based on Spatio-Temporal Resolution Adaptation", *IEEE Trans. on CSVT*.

[Zhang *et al.*, 2021] "ViSTRA2: Video coding using spatial resolution and effective bit depth adaptation", *Signal Processing: Image Communication*.

Perceptual Quality Comparison: ParkRunning

[VIDEO] **Topleft**: Reconstructed video of sequence *ParkRunning* for the HM anchor. **Bottomleft**: The corresponding video for ViSTRA-HM at the same bitrate. **Middle**: The video for the enlarged block of the top left video. **Right**: The video for the enlarged block of the bottom left video (the same location).

AI-based Coding Tools: links to Existing Standards

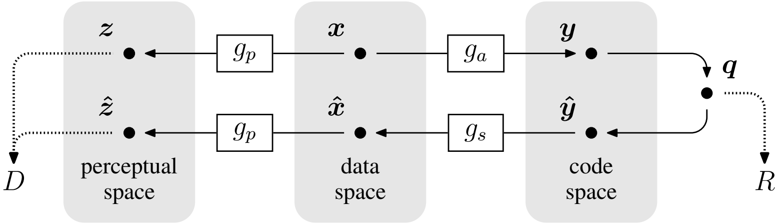
- MPEG JVET are developing an AI-optimised video codec, **NNVC**, on top of VVC VTM 11.
- NNVC (v-10.0) offers up to **14%** coding gains over VTM, but with a high decoder complexity increase.
- AOM is also considering neural network based solutions (complexity lower than **2k MACs/pixels**).
- One of the best AVM tools offers a **3.9%** BD-rate saving in PSNR-Y, with a complexity of **1500 MACs/pixel**.
- Most of these tools are based on **post-processing** (or in-loop filtering) and/or **super-resolution**.
- The **trade-off** between complexity and performance remains a challenge for this type of solution.

[Galpin *et al.*, 2024] "JVET AHG report: NNVC software development AhG14", JVET-AJ0014.

[Joshi *et al.*, 2023] "Switchable CNNs for in-loop restoration and super-resolution for AV2", SPIE2023.

Learned Video Compression via End-to-end Optimisation

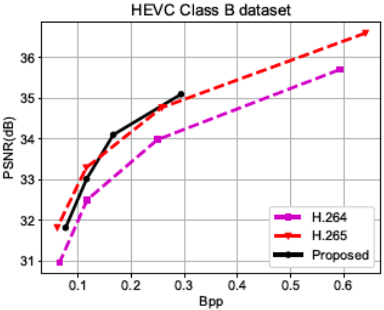
- Traditional codec tool enhancements remains the dominant approach currently.
- However, inspired by the success of end-to-end learned image compression [Ballé *et al.*, 2017, 2018]. significant advances in end-to-end learned **video** codecs are emerging, that are **holistically optimisable**.

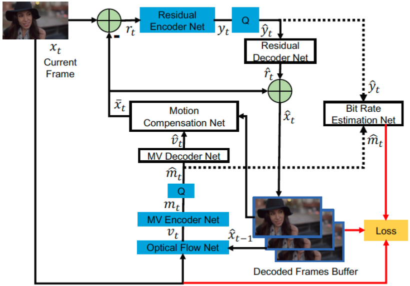


[Ballé *et al.*, 2017] "End-to-end optimized image compression", International Conference on Learning Representations.

End-to-End Learned Video Codecs

- **DVC** [Lu *et al.*, 2019] was the first end-to-end deep video compression model.
- Replaces the conventional video coding framework with several **neural networks**.
- Achieves a performance similar to **x265 (*veryfast* preset)**.

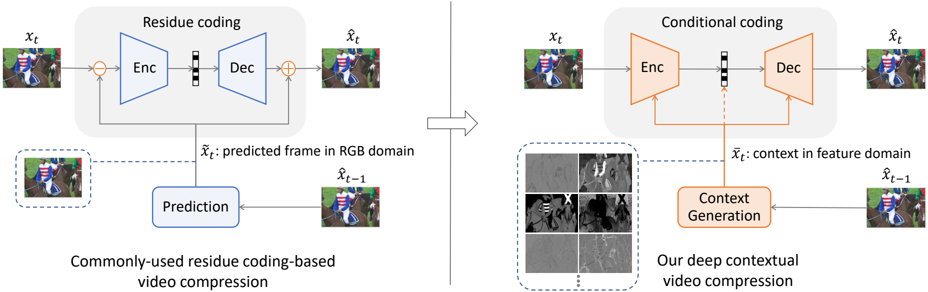




Source of figures: [Lu *et al.*, 2019] Lu et al., "DVC: An end-to-end deep video compression framework", IEEE/CVF CVPR, 2019.

Deep Contextual Video Compression (DCVC)

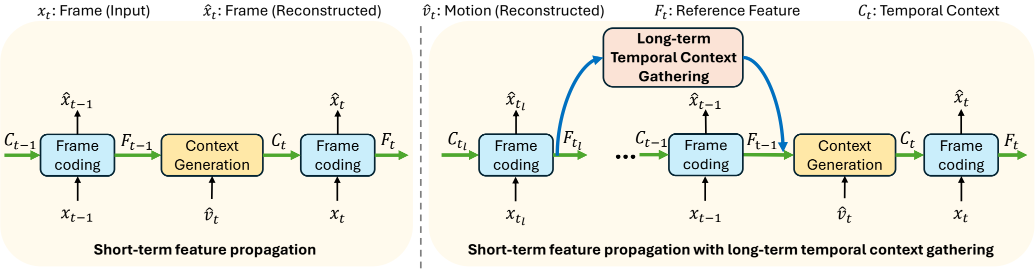
- A series of neural video codecs that offer similar RQ performance to standard video codecs.
- Shift from a residue coding - to a **conditional coding-based** framework.



Source of figure: [Li *et al.*, 2021] "Deep contextual video compression", NeurIPS.

DCVC Codecs

- **DCVC-DC**: diverse contexts [Li *et al.*, 2023] - better than ECM LD (RGB).
- **DCVC-FM**: feature modulation [Li *et al.*, 2024] - better than ECM LD (RGB and YUV) .
- **DCVC-LCG**: long-term temporal context gathering [Qi *et al.*, 2024] - **11.3%** better than ECM LD (YUV).



(a) Previous methods (b) Our DCVC-LCG

Source of figure: [Qi *et al.*, 2024] “Long-term Temporal Context Gathering for Neural Video Compression”, ECCV.

Limitations of AI-based Video Coding Methods

Coding performance

- Reported results for (most) learned video codecs are not based on **standard CTCs**.
- Conventional video codecs (RA mode) still lead in performance.

Complexity Issues

- Pre-trained generic models typically require **large model capacity**.
- This leads to significantly **increased** (decoding) **complexity** and **large model sizes**.

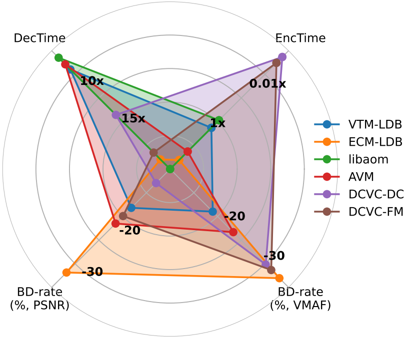
Non-standard pipelines

- Existing neural video codecs adopt **diverse pipelines/network architectures**.
- **Compatibility** is essential in video coding (standardisation) and **convergence** is required.

Conventional and Learned Video Codecs - Benchmarking

- **Configuration**: low delay, YUV420 and JVET/AOM test sequences.
- **Hardware**: PC with single CPU (Intel i7-12700) and GPU (NVIDIA 3090).

	BD-rate (PSNR)	BD-rate (VMAF)	Encoding time	Decoding time
libaom	0%	0%	1×	1×
DCVC-DC	-11.2%	-31.1%	0.008×	13.180×
VTM-LDP	-13.4%	-18.2%	1.032×	2.723×
VTM-LDB	-19.2%	-22.5%	1.502×	2.675×
DCVC-FM	-20.8%	-33.1%	0.012×	20.735×
AVM	-21.4%	-25.6%	12.301×	2.246×
ECM-LDP	-29.0%	-30.4%	10.704×	21.394×
ECM-LDB	-33.9%	-34.2%	16.106×	21.725×



[Teng *et al.*, 2024] “Benchmarking Conventional and Learned Video Codecs with a Low-Delay Configuration.”, IEEE VCIP.

When will AI-based Methods be Acceptable?

Architecture and Complexity reduction

- Significant complexity reduction, especially at the decoder.
- Performance should be maintained with low-complexity models.
- Architectural convergence for standardisation.

Coding gains

- More significant and consistent coding gains over best standard codecs (ECM/AVM).

Rate quality optimisation

- Exploitation of perceptual redundancy during coding.
- Better quality assessment methods and loss functions.

Outline

Video Compression - pre AI

AI-based Video Compression

Reducing Complexity

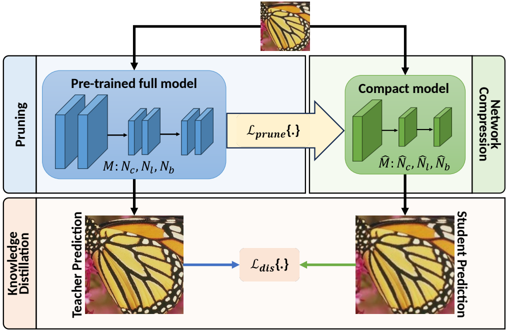
Motion Models

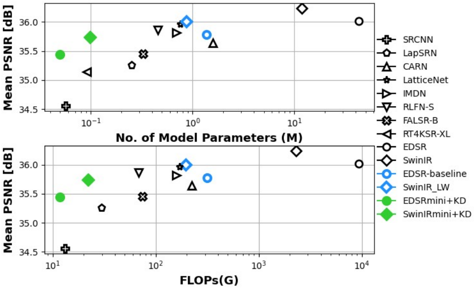
Representation Models

Conclusion

Model Compression and Knowledge Distillation - e.g. ISR

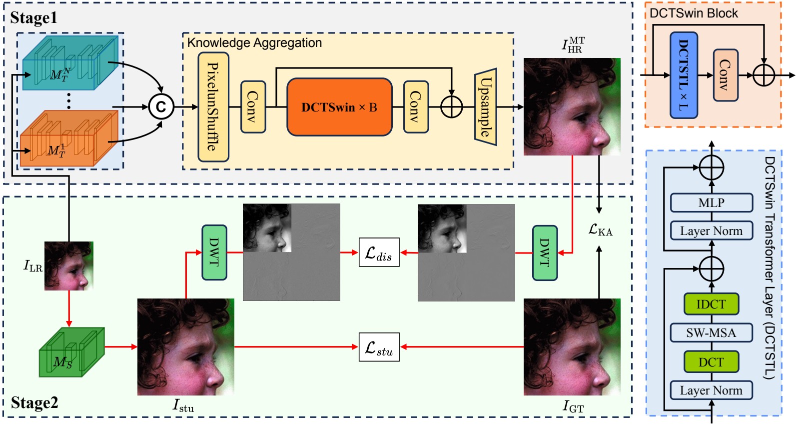
- Model complexity can be significantly reduced by **model compression**.
- Model performance after compression can be further improved through **knowledge distillation**.





[Jiang *et al.*, 2024c] "Compressing deep image super-resolution models", PCS.

MTKD: Multi-Teacher Knowledge Distillation

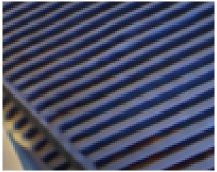


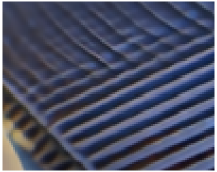
[Jiang *et al.*, 2024a] "MTKD: multi-teacher knowledge distillation for image super-resolution", ECCV.

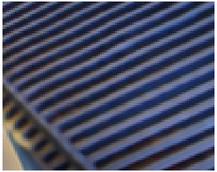
MTKD: Results

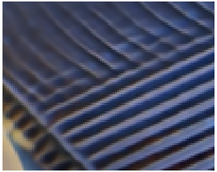
SwinIR vs SwinIR_lightweight [Liang *et al.*, 2021] (90% complexity reduction)

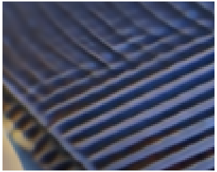

HR
PSNR/SSIM

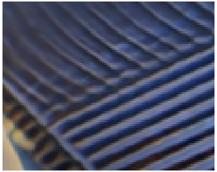

Full
24.81/.7928

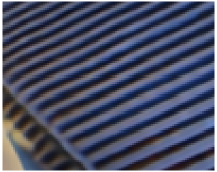

Compact
24.20/.7542


MT
24.93/.7865


KD
24.19/.7533

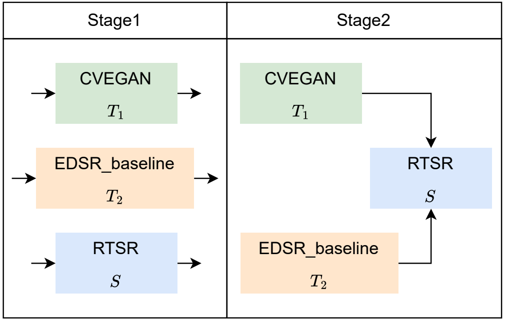

AT
24.19/.7530

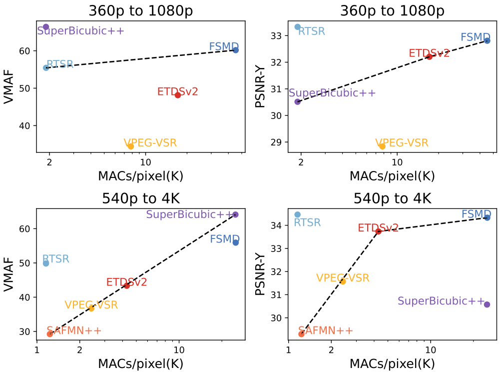

FAKD
24.14/.7526


MTKD (Ours)
24.34/.7622

RTSR: Real-Time SR for Compressed Content

Extend from image super-resolution to video compression (AV1).

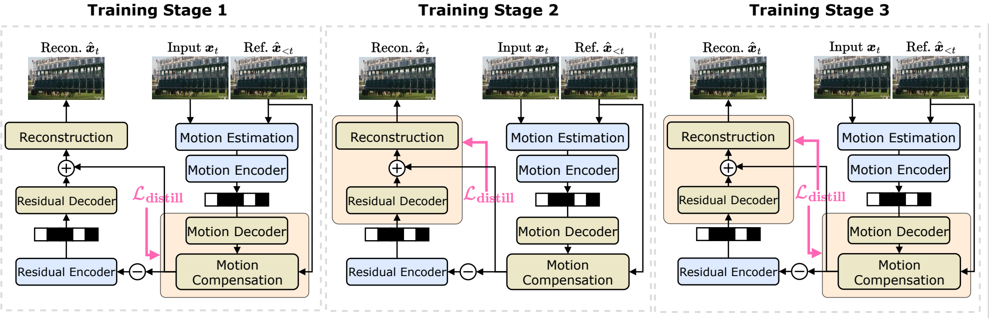




[Jiang *et al.*, 2024b] "RTSR: A Real-Time Super-Resolution Model for AV1 Compressed Content", submitted to ISCAS 2025.

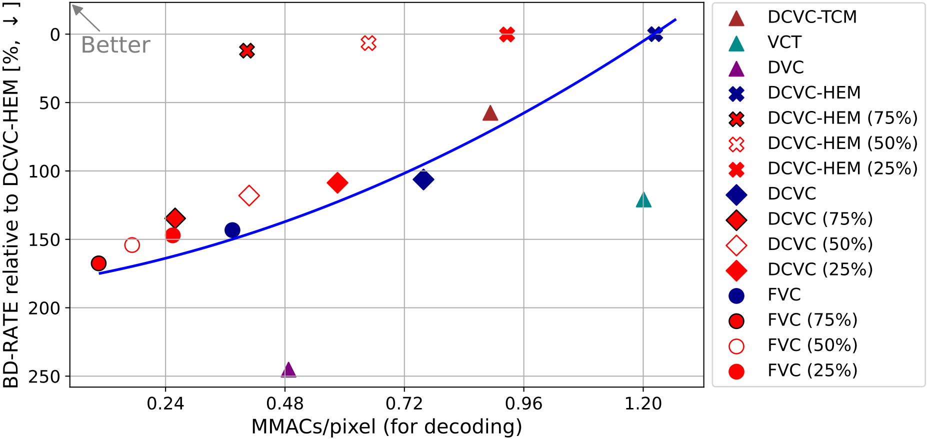
End-to-end Learned Video Codecs: Complexity Reduction

1. The **multi-stage optimisation** of learnt video codecs vs the **global pruning objectives**.
2. Split the distillation of sub-modules into **multi-stages** to regularise the student model.



[Peng *et al.*, 2024] "Accelerating learnt video codecs with gradient decay and layer-wise distillation.", PCS.

Evaluation Results - Complexity vs Performance



Outline

Video Compression - pre AI

AI-based Video Compression

Reducing Complexity

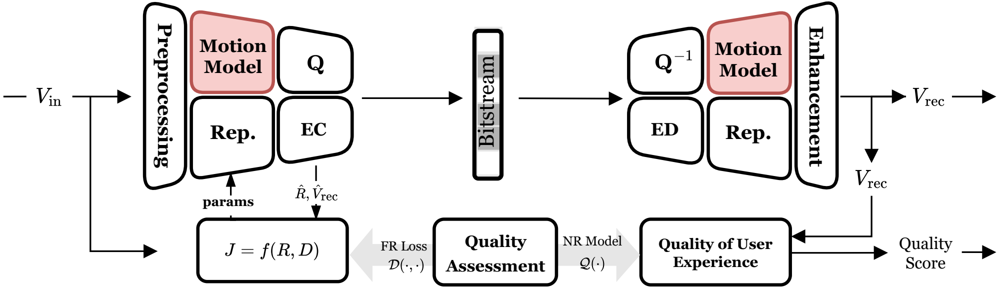
Motion Models

Representation Models

Conclusion

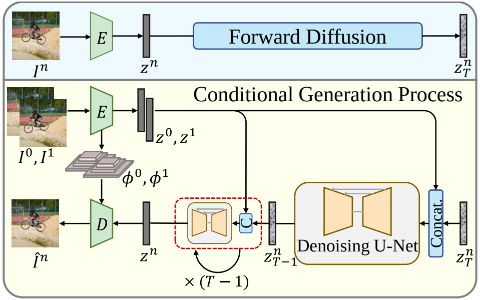
Motion Models in Video Compression

- Accurate **motion estimation** is key in exploiting the temporal redundancy within videos.
- **Video frame interpolation** techniques offer potential solutions for improved motion modelling.



Perceptually-oriented VFI: LDMVFI

- **L1/L2/VGG loss** does not correlate with VFI perceptual quality.
- **Diffusion models** have shown remarkable performance in generating perceptually-optimised images.
- We tailor **latent diffusion models** for VFI to achieve superior perceptual quality.

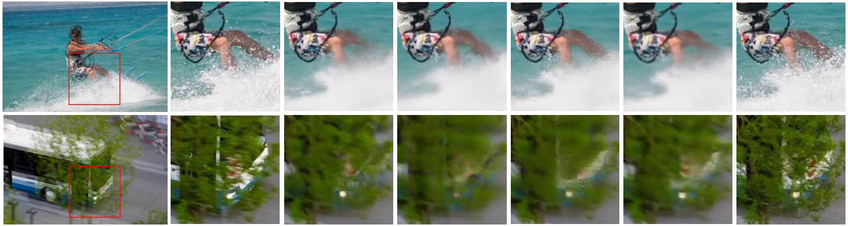


[Danier *et al.*, 2024] "LDMVFI: Video Frame Interpolation with Latent Diffusion Models", AAAI.

LDMVFI: Performance

	Middlebury			UCF-101			DAVIS			VFITex			RT (sec)	#P (M)
	LPIPS↓	FloLPIPS↓	FID↓	LPIPS↓	FloLPIPS↓	FID↓	LPIPS↓	FloLPIPS↓	FID↓	LPIPS↓	FloLPIPS↓	FID↓		
BMBC	0.023	**0.037**	12.974	0.034	0.045	33.171	0.125	0.185	15.354	0.220	0.282	50.393	0.51	11.0
AdaCoF	0.031	0.052	15.633	0.034	0.046	32.783	0.148	0.198	17.194	0.204	0.273	42.255	0.01	21.8
IFRNet	0.020	0.039	12.256	0.032	0.044	28.803	0.114	0.170	14.227	0.200	0.273	42.266	0.02	5.0
VFIformer	0.031	0.065	15.634	0.039	0.051	34.112	0.191	0.242	21.702	OOM	OOM	OOM	1.74	5.0
ST-MFNet	N/A	N/A	N/A	0.036	0.049	34.475	0.125	0.181	15.626	0.216	0.276	41.971	0.14	21.0
FLAVR	N/A	N/A	N/A	0.035	0.046	31.449	0.209	0.248	22.663	0.234	0.295	56.690	0.02	42.1
MCVD	0.123	0.138	41.053	0.155	0.169	102.054	0.247	0.293	28.002	OOM	OOM	OOM	52.55	27.3
LDMVFI	**0.019**	0.044	16.167	**0.026**	**0.035**	**26.301**	**0.107**	**0.153**	**12.554**	**0.150**	**0.207**	**32.316**	8.48	439.0

[Video] Visual comparison between different VFI models.



Overlayed Inputs GT BMBC VFIformer IFRNet ST-MFNet LDMVFI (Ours)

Outline

Video Compression - pre AI

AI-based Video Compression

Reducing Complexity

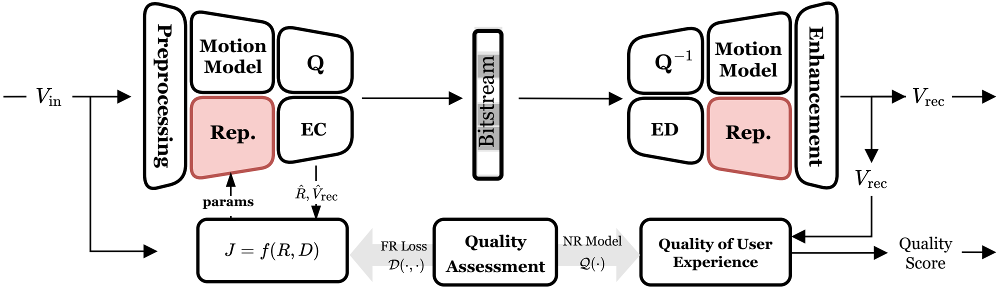
Motion Models

Representation Models

Conclusion

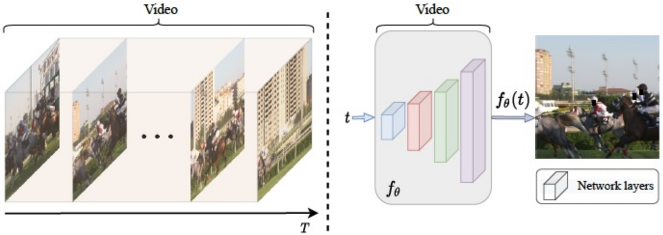
Advanced Representation Models

- Existing end-to-end learned video codecs suffer from **high computational complexity**.
- **Amortised inference**: hyperparameters are fixed and shared across diverse contents.
- Therefore **more sophisticated architectures** are required.



Neural Representation for Videos (NeRV)

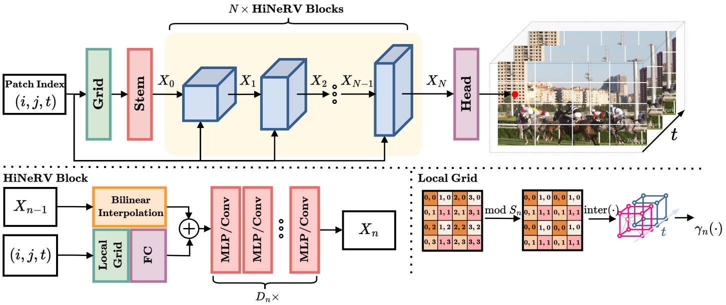
- Implicit Neural Representations offer a promising solution, **overfitting** the content during learning.
- Neural Radiance Field (NeRF): $f : f(x, y, z, \theta, \phi) = (r, g, b, \sigma)$.
- Neural Representation for Videos (NeRV): $f : f(x, y, t) = (r, g, b)$.
- NeRV-based video codecs can offer very **fast decoding speed**.
- However, existing INR models (e.g., NeRV [Chen *et al.*, 2021] and HNeRV [Chen *et al.*, 2023]) are **not competitive** against standard or other E2E learned codecs.



Souce of figure: [Chen *et al.*, 2021] "NeRV: Neural representations for videos", NeurIPS.

HiNeRV: Hierarchical Encoding-based Neural Representation

- A new upsampling layer with bilinear interpolation and **hierarchical encoding** of feature grids.
- A **unified representation** of frame- and patch-wise INR by adding padding for acceleration.
- A **refined training pipeline**, with pruning- & quantisation-aware fine-tuning.



[Kwan *et al.*, 2023] "Video Compression with Hierarchical Encoding-based Neural Representation", NeurIPS.

HiNeRV: Performance

- HiNeRV is the first INR-based codec that outperforms HEVC x265 (*veryslow*).
- It also outperforms existing NeRV-based video codecs with up to **70%** bit rate-savings.
- and offers fast decoding speed - up to **35FPS**.

Dataset	Metric	x265 (*veryslow*)	HM (*RA*)	DCVC	DCVC-HEM	VCT	NeRV	HNeRV
UVG	PSNR	-38.66%	7.54%	-43.44%	25.23%	-34.28%	-74.12%	-72.29%
	MS-SSIM	-62.70%	-41.41%	-34.50%	49.03%	-23.69%	-73.76%	-83.86%
MCL-JCV	PSNR	-23.39%	31.09%	-24.59%	35.83%	-17.03%	-80.19%	-66.56%
	MS-SSIM	-44.12%	-2.65%	-17.32%	80.73%	12.10%	-82.28%	-79.42%

GT

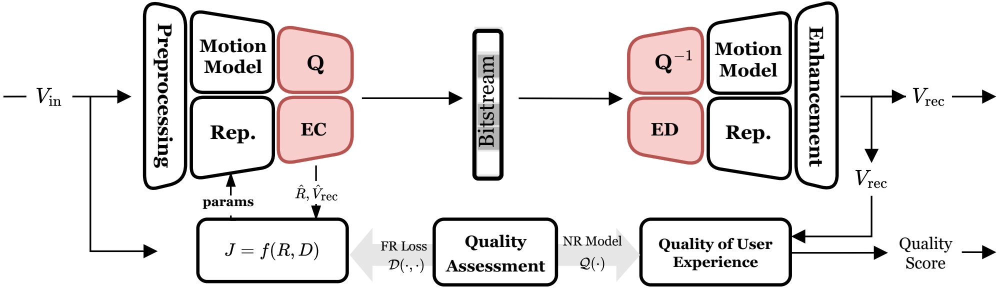
NeRV
31.4dB PSNR@0.099bpp

HNeRV
31.4dB PSNR@0.101bpp

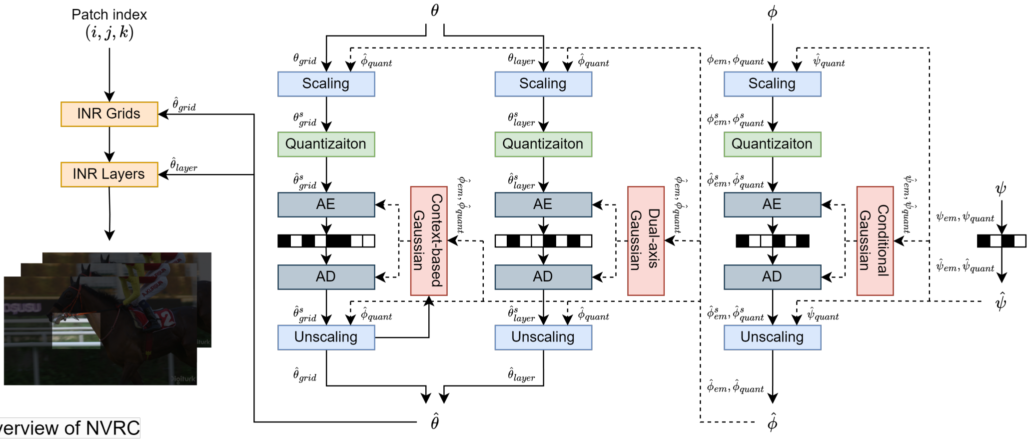
HiNeRV (ours)
36.6dB PSNR@0.051bpp

Improving Quantisation and Entropy Coding

- **Quantization** creates lossy representations of input videos.
- Accurate **entropy modelling** is also key to high compression ratios.



NVRC: Neural Video Representation Compression

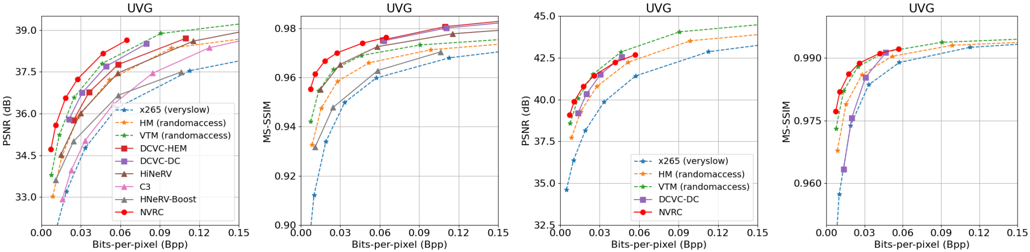


Overview of NVRC

[Kwan *et al.*, 2024a] “NVRC: Neural Video Representation Compression”, NeurIPS.

NVRC: Performance

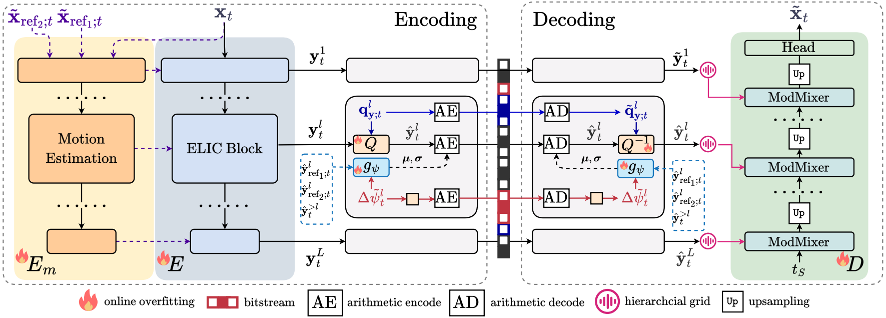
Color Space	Metric	x265 (*veryslow*)	HM (*RA*)	VTM (*RA*)	DCVC-HEM	DCVC-DC	HiNeRV	C3	HNeRV-Boost
RGB 4:4:4	PSNR	-74.02%	-51.00%	-24.34%	-41.30%	-32.05%	-50.73%	-67.93%	-66.78%
	MS-SSIM	-80.79%	-67.61%	-50.08%	-7.91%	-12.58%	-44.69%	-	-78.21%
YUV 4:2:0	PSNR	-62.71%	-34.83%	-1.03%	-	-62.28%	-	-	-
	MS-SSIM	-59.49%	-38.45%	-15.38%	-	-70.23%	-	-	-



[VIDEO] Visual Comparison between HM and NVRC.

PNVC: Towards Practical Neural Video Compression

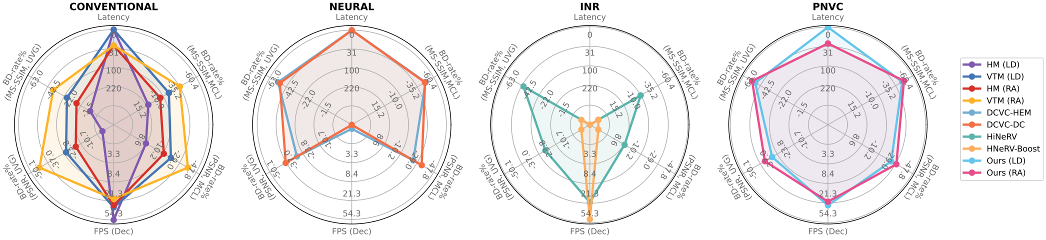
- INR-based video codecs typically represent an entire video or a dataset with a single monolithic model.
- This requires processing a large number of frames in each encoding, resulting in a **high system latency**.
- **PNVC**: a practical neural video coding framework, enabling **flexible coding configurations** (LD and RA).



[Gao *et al.*, 2025] "PNVC: Towards Practical INR-based Video Compression", AAAI.

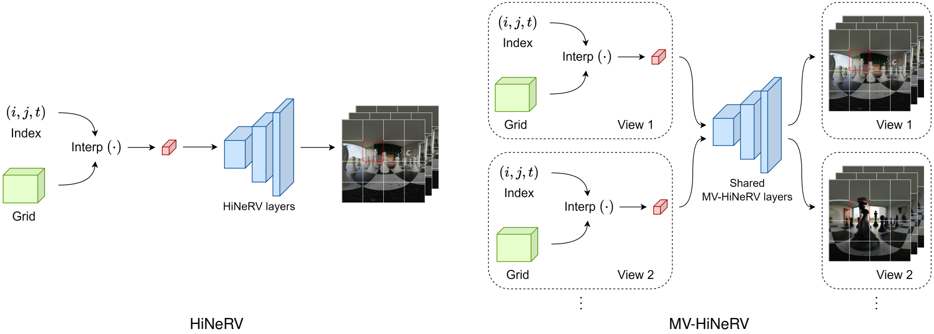
PNVC: Performance

- **Well-rounded** performance across multiple dimensions.
- **5%+** gain over VTM (LD) in PSNR and MS-SSIM.
- **10%+** gain over HiNeRV in PSNR and MS-SSIM.
- **20+FPS** decoding speed for HD (1080P).
- Flexible coding/delay configurations (**LD** and **RA**).



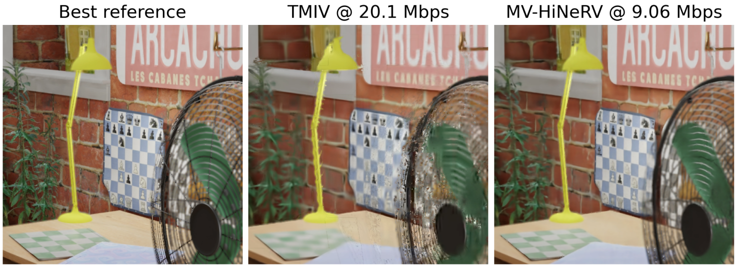
MV-HiNeRV - Extending HiNeRV to Immersive Videos

- **MV-HiNeRV** [Kwan *et al.*, 2024b] extends HiNeRV to the compression of **immersive/volumetric videos**.
- It learns hierarchical feature grids **per view**, and **shares** the learned network parameters among all views.
- This enables the model to effectively exploit the spatio-temporal and the **inter-view redundancy**.
- MV-HiNeRV has achieved significant coding gains (up to **72.33%**) over MPEG TMIV (based on VVenc).



MV-HiNeRV: Performance

BD-rate (%)	B02	D01	E01	J02	J04	W01	Overall
PSNR	-17.60	-65.93	-36.59	-59.96	-80.03	-35.48	-49.27
IV-PSNR	-38.11	-61.08	-6.28	-70.21	-72.33	-33.50	-46.92



[Kwan *et al.*, 2024b] "Immersive Video Compression using Implicit Neural Representations", PCS.

Outline

Video Compression - pre AI

AI-based Video Compression

Reducing Complexity

Motion Models

Representation Models

Conclusion

Summary and Future Work

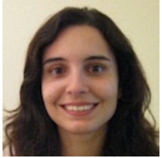
Learning-based video coding: Hope or Hype?

- **Deep learning** has made important contributions to video compression and quality assessment.
- **But** significant issues remain include coding **performance**, **complexity**, and non-standard **pipelines**.
- **Generative methods**: enable super-resolution and motion interpolation tools for near term advances.
- **INR-based frameworks**: potential for the best trade-off between complexity, performance and practicality.
- **Complexity reduction**: enabled by model compression and knowledge distillation.

Future work

- **Performance**: demonstrate significant coding gains over ECM/AVM with lower model complexity.
- **Evaluation**: new metrics and benchmarking methods to compare techniques with varying performance characteristics, consistency and artefacts.
- **Convergence**: stable architectures to drive investment in standards and hardware.
- **Compatibility**: with low cost integrated hardware: NPU and TPU acceleration.
- **Datasets**: relevant, diverse and extensive,

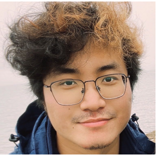
Contributors


Mariana Afonso


Duolikun Danier


Chen Feng


Ge Gao


Yuxuan Jiang

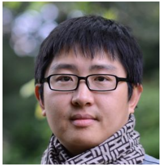

Ho Man Kwan


Di Ma


Jakub Nawala


Jasmine Peng


Siyue Teng


Aaron Zhang

Funders and Collaborators

prime video

BBC

BT

bvi BRISTOL VISION INSTITUTE

CONDENSE

CSC SCHOLARSHIP A WAY TO YOUR DREAMS

EPSRC Engineering and Physical Sciences Research Council

PROVISION Perceptually Optimised Video Compression

MyWorld

NETFLIX

Tencent 腾讯

UK Research and Innovation

Visionular

Thank you!

Q & A

References I

Mariana Afonso, Fan Zhang, and David R Bull. Video compression based on spatio-temporal resolution adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(1):275–280, 2018.

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 221–231, 2019.

Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018.

Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser Nam Lim, and Abhinav Shrivastava. Nerv: Neural representations for videos. *Advances in Neural Information Processing Systems*, 34:21557–21568, 2021.

Hao Chen, Matthew Gwilliam, Ser-Nam Lim, and Abhinav Shrivastava. Hnerv: A hybrid neural representation for videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10270–10279, 2023.

Duolikun Danier, Fan Zhang, and David Bull. Ldmvfi: Video frame interpolation with ålatent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1472–1480, 2024.

F. Galpin, R. Chang, Y. Li, Y. Li, M. Santamaria, J. N. Shingala, and Z. Xie. JVET AHG report: NNVC software development AhG14. In *JVET-AJ0014*, 2024.

Ge Gao, Ho Man Kwan, Fan Zhang, and David Bull. Pnvc: Towards practical inr-based video compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Yuxuan Jiang, Chen Feng, Fan Zhang, and David Bull. Mtkd: Multi-teacher knowledge distillation for image super-resolution. In *ECCV*, 2024.

References II

Yuxuan Jiang, Jakub Nawała, Chen Feng, Fan Zhang, Xiaoqing Zhu, Joel Sole, and David Bull. Rtsr: A real-time super-resolution model for av1 compressed content. *arXiv preprint arXiv:2411.13362*, 2024.

Yuxuan Jiang, Jakub Nawała, Fan Zhang, and David Bull. Compressing deep image super-resolution models. In *2024 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2024.

Urvang Joshi, Yue Chen, Innfarn Yoo, Shan Li, Feng Yang, and Debargha Mukherjee. Switchable cnns for in-loop restoration and super-resolution for av2. In *Applications of Digital Image Processing XLVI*, volume 12674, pages 121–130. SPIE, 2023.

Ho Man Kwan, Ge Gao, Fan Zhang, Andrew Gower, and David Bull. Hinerv: Video compression with hierarchical encoding-based neural representation. *Advances in Neural Information Processing Systems*, 36, 2023.

Ho Man Kwan, Ge Gao, Fan Zhang, Andrew Gower, and David Bull. Nvrc: Neural video representation compression. *Advances in Neural Information Processing Systems*, 2024.

Ho Man Kwan, Fan Zhang, Andrew Gower, and David Bull. Immersive video compression using implicit neural representations. *arXiv preprint arXiv:2402.01596*, 2024.

Tianyi Li, Mai Xu, Ce Zhu, Ren Yang, Zulin Wang, and Zhenyu Guan. A deep learning approach for multi-frame in-loop filter of hevc. *IEEE Transactions on Image Processing*, 28(11):5663–5678, 2019.

Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. *Advances in Neural Information Processing Systems*, 34, 2021.

Jiahao Li, Bin Li, and Yan Lu. Neural video compression with diverse contexts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, Canada, June 18-22, 2023*, 2023.

Jiahao Li, Bin Li, and Yan Lu. Neural video compression with feature modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 17-21, 2024*, 2024.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

References III

Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2507–2515, 2017.

Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. DVC: An end-to-end deep video compression framework. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11006–11015, 2019.

Di Ma, Fan Zhang, and David R Bull. Bvi-dvc: A training database for deep video compression. *IEEE Transactions on Multimedia*, 24:3847–3858, 2021.

Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

Jakub Nawała, Yuxuan Jiang, Fan Zhang, Xiaoqing Zhu, Joel Sole, and David Bull. Bvi-aom: A new training dataset for deep video compression optimization. In *IEEE Visual Communications and Image Processing*, 2024.

Tung Nguyen and Detlev Marpe. Compression efficiency analysis of av1, vvc, and hevc for random access applications. *APSIPA Transactions on Signal and Information Processing*, 10:e11, 2021.

Tianhao Peng, Ge Gao, Heming Sun, Fan Zhang, and David Bull. Accelerating learnt video codecs with gradient decay and layer-wise distillation. In *2024 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2024.

Linfeng Qi, Zhaoyang Jia, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Long-term temporal context gathering for neural video compression. In *ECCV*, 2024.

Vadim Seregin, Jie Chen, Roman Chernyak, Fabrice Le Léannec, and Kai Zhang. JVET AHG report: ECM software development (AHG6). In *JVET-AI0006*, 2024.

References IV

Siyue Teng, Yuxuan Jiang, Ge Gao, Fan Zhang, Thomas Davis, Zoe Liu, and David Bull. Benchmarking conventional and learned video codecs with a low-delay configuration. In *IEEE Visual Communications and Image Processing*, 2024.

Haiqiang Wang, Ioannis Katsavounidis, Jiantong Zhou, Jeonghoon Park, Shawmin Lei, Xin Zhou, Man-On Pun, Xin Jin, Ronggang Wang, Xu Wang, et al. Videoset: A large-scale compressed video quality dataset based on jnd measurement. *Journal of Visual Communication and Image Representation*, 46:292–302, 2017.

Stefan Winkler. Analysis of public image and video databases for quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):616–625, 2012.

Fan Zhang and David R Bull. A parametric framework for video compression using region-based texture models. *IEEE Journal of Selected Topics in Signal Processing*, 5(7):1378–1392, 2011.

Fan Zhang, Mariana Afonso, and David R Bull. ViSTRA2: Video coding using spatial resolution and effective bit depth adaptation. *Signal Processing: Image Communication*, 97:116355, 2021.