Perceptual Video Compression with Neural Wrapping

Muhammad Umar Karim Khan, Aaron Chadha, Mohammad Ashraful Anam, Yiannis Andreopoulos
Sony Interactive Entertainment
{Umar . Khan, Aaron . Chadha, Russell . Anam, Yiannis . Andreopoulos}@sony.com

Abstract

Standard video codecs are rate-distortion optimization machines, where distortion is typically quantified using PSNR versus the source. However, it is now widely accepted that increasing PSNR does not necessarily translate to better visual quality. In this paper, a better balance between perception and fidelity is targeted, in order to provide for significant rate savings over state-of-the-art standards-based video codecs. Specifically, pre- and post-processing neural networks are proposed that enhance the coding efficiency of standard video codecs when benchmarked with an array of well-established perceptual quality scores. These "neural wrapper" elements are end-to-end trained with a neural codec module serving as a differentiable proxy for standard video codecs. The codec proxy is jointly optimized with the pre- and post components via a novel two-phase pretraining strategy and end-to-end iterative refinement with stop-gradient. This allows the neural pre- and postprocessor to learn to embed, remove and recover information in a codec-aware manner, thus improving its rate-quality performance. A single neural-wrapper model is thereby established and used for the entire rate-quality curve without needing any downscaling or upscaling. The trained model is tested with the AV1 and VVC standard codecs via an array of well-established objective quality scores (SSIM, MS-SSIM, VMAF, AVQT), as well as mean opinion scores (MOS) derived from ITU-T P.910 subjective testing. Experimental results show that the proposed approach improves all quality scores, with -18.5% average Bjontegaard Delta-rate (BD-rate) saving over all objective scores and MOS improvement over both standard codecs. This illustrates the significant potential of neural wrapper components over standards-based video coding.

1. Introduction

It is now widely accepted that that signal-to-noise (SNR) ratio is a poor indicator of visual quality in video coding [6, 19, 26, 42]. Instead, quality scores that include elements of human perception [42], perceptual modelling of encoding artifacts [51], as well as viewing setup awareness [4, 65] have emerged as strong contenders for the best means to objectively characterize visual quality. This has also led to the research community now moving away from SNR-optimization [26] in favor of structural similarity (SSIM [65]), video multimethod assessment fusion (VMAF) and Apple's advanced video quality tool (AVQT) optimization [4, 42]. However, all current perceptual optimization approaches in standards-based video encoders have one or more of the following detriments:

- they require multiple encoding passes or in-loop implementation within a specific encoder;
- they only optimize for a single quality scoring method like VMAF or SSIM and are shown to be detrimental in other quality scores;
- they comprise hand-crafted (shallow) models of low-level human perception and fail to encapsulate several characteristics of more advanced quality scoring methods like VMAF or AVQT in a data-driven and learnable manner;
- their Bjontegaard Delta-rate (BD-rate) improvement on well-established quality scores like SSIM, VMAF and AVQT is in the order of a few percentile points.

A neural codec "wrapper" is proposed in this paper as the means to augment standard video codecs and optimize them for perceptual quality. Including neural pre- and postprocessing in conventional encoding pipelines is shown to clearly outperform the equivalent codecs without these components. This allows for conventional streaming ecosystems to continue operating "as-is" and does not break standards. That is, when neural components cannot be supported before encoding or after decoding, the system can just disable them with appropriate adaptation, without jeopardizing video streaming. The key contributions of the paper are listed as follows.

  - A novel pretraining regime for modelling a conventional codec with a neural codec model is proposed, which uses the implicit encoder-decoder structure of the neural codec to align both video quality and rate.
  - An end-to-end training regime for jointly training pre and postprocessors is proposed. Starting with a pretrained neural codec model, it is proposed to iteratively update the codec model. This approach allows for more accurate gradients to be backpropagated to the prepro-

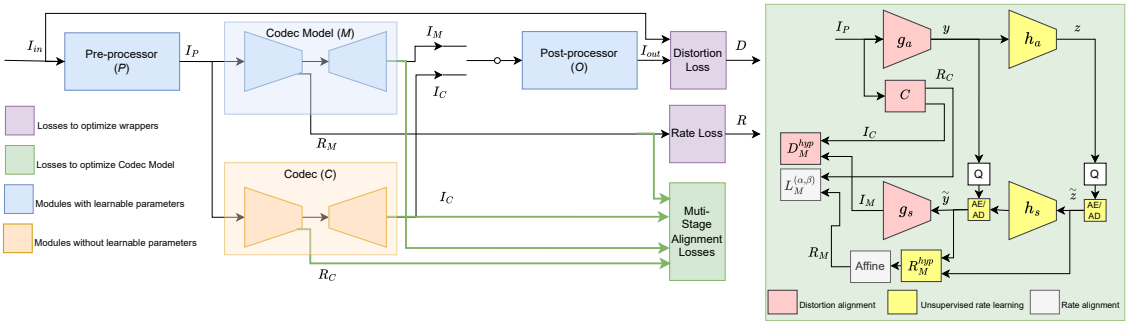


Figure 1. Left: Functional block diagram of the multi-stage training data flow. Right: Components of codec model M that are used as a proxy for the target codec C. *AE/AD* denotes arithmetic encoding followed by arithmetic decoding and Q denotes quantization. See Sec. 3 for details.

cessor.

– The trained pipeline is evaluated on gaming sequences and significant BD-rate gains versus AV1 and VVC encoding are demonstrated with just a single pre- and postprocessor model. In addition, the proposed neural wrapper is shown to outperform a state-of-the-art neural encoder [35], as well as the state-of-the-art in deep perceptual preprocessing proposed recently [11].

A functional block diagram for the proposed system is shown in Fig. 1. Example visuals produced by the proposed method are shown in Fig. 4, demonstrating that the proposed neural wrapper framework allows for better detail preservation than the underlying standard codec.

2. Related Work

2.1. Neural Video Codecs

Neural video codecs optimize neural networks for rate and distortion by exploiting redundancies in video data. Most notable are residual coding methods; Lu *et al.* [46] introduce deep video codec, which adopts neural networks within the residual coding pipeline of a traditional codec. Optical flow is employed as a proxy for motion vectors in order to warp the reference frame and generate a residual frame for encoding. Numerous other methods have been proposed based on residual coding [44, 49, 53]. Scale-space flow estimation is proposed as an improvement over traditional optical flow by Agustsson *et al.* [2], namely for better handling of complex motion in the scene. Beyond residual coding methods, conditional coding methods learn feature contexts implicitly from the encoded context [23, 33, 34, 45, 47, 56]. Li *et al.* [35] propose using diverse contexts in spatial and temporal directions to perform video compression and outperform standard video codecs when the latter are used in low-delay mode. They improve the performance and bitrate range in [36]. Recent work has additionally focused on reducing the complexity of neural codecs. For example, Van Rozendaal *et al.* [64] deploy an architecture similar to Agustsson *et al.* [2] on a smart phone and achieve real time performance at 1080p. Hu *et al.* [25] adopt slimmable networks for variable complexity.

2.2. Preprocessors

While postprocessing is a more mature line of research, there has been a prevalence of work in preprocessing, both on image and video modalities. In the image domain, Strumpler *et al.* [58] propose to modify the JPEG encoder with an attention network and a learnable quantization table, in order to modulate the source frame and provide for additional rate savings over the JPEG encoder alone. Conversely, Talebi *et al.* [59] avoid adapting the JPEG encoder by prepended the standard encoder with a preprocessor for deployment with a standard JPEG encoder and decoder. Klopp *et al.* [31] train a preprocessor with a surrogate neural codec model, which is trained to model a target image codec. In the video domain, Bourtsoulatze *et al.* [9] propose a precoding network for downscaling with a online mode selection algorithm, in order to alter the precoding based on scene characteristics. Conversely, Chadha *et al.* [11] propose an entirely offline training mechanism by first modelling the core components of a video codec in a differentiable manner. A preprocessor is optimized on the codec model, in order to provide further rate-quality savings over the target codec.

2.3. Postprocessors

Numerous deep learning methods have been proposed for compressed video denoising, video super-resolution or both tasks. Single frame methods for denoising compressed videos have been proposed in [71, 72]. Multi-frame methods have provided superior gains in metrics compared to single frame methods. MFQE [73] detects frames with higher quality and uses it as a reference to

improve neighboring frames. TSAN [69] uses a temporal deformable alignment and multi-scale fusion for the task of video restoration. STDF [16] uses spatio-temporal deformable convolution to aggregate temporal context rather than using optical flow for frame warping. A common approach for video super-resolution is to align neighboring frames and transfer information across frames [60, 61, 66, 70]. On the contrary, other methods choose non-local modules or 3D convolutions to gather temporal context [29, 30, 38, 74]. Some methods recurrently propagate history in the forward direction [17, 28, 55]. Recent methods such as BasicVSR [12], GOVSR [75] and BasicVSR++ [13] propagate history in forward and backward directions, such that both past and future frames are used to super-resolve the current frame. Architectures developed for video super-resolution can easily be adapted for video denoising. Few authors have focused on the task of jointly denoising and upscaling videos. FTVSR [50] uses a frequency transformer which performs self-attention in space, time and frequency to recover details. RealBasicVSR [14] considers sensor noise, motion blur and compression artifacts with super-resolution. CIAF [76] uses motion-vectors from the video to approximate optical flow. CAVSR [67] adapts the super-resolution to the compression level of the input video. COMISR [39] uses an architecture based on BasicVSR [12] to perform joint denoising and upscaling.

2.4. Pre- and Postprocessors with Traditional Codecs

Recent methods have also explored joint training of pre and postprocessors as neural wrappers. In the image domain, recent work [20, 21] emulates the core blocks of the JPEG encoder and decoder, and uses it to train pre and postprocessors for rate-distortion savings. Similar sandwich approach for videos has been proposed in [24]. Son *et al.* [57] propose distinct neural networks to model the distortion and rate behavior of a standard codec, which is thus sandwiched by pre- and postprocessor networks for rate-distortion optimization. However, these methods have no target codec awareness beyond the implicit intra/inter coding structure. Either the codec proxy is not learned (e.g. derived from differentiable JPEG/DCT as in [24]), only partially learned via rate-distortion optimization (e.g. the loop filter as in [27]) or only operates in intra mode (as in [57]). In this paper, the proposed framework is directly conditioned on the target conventional codec. Tian *et al.* [62] target additional rate savings with learnable rescaling. However, this method has no rate control or alignment outside of the scaling factor. Conversely, the framework proposed in this paper is jointly conditioned on both the target codec rate and distortion and is thus able to offer rate savings even when encoding at full resolution. Additionally, whereas recent work restricts evaluation to a limited range of the rate-distortion space, at high bitrates [61] or substantially lower bitrates and resolutions [22] and on older encoding standards (AVC or HEVC), the proposed method is evaluated against state-of-the-art AV1 and VVC encoding, i.e.: *(i)* the entire rate-quality region of practical significance is considered (VMAF values between 40-96 and bitrates from 100kbps to 15mbps); *(ii)* slow encoding presets are used (preset=4 for SVT-AV1 and preset=slow for VVenC), *without* limiting the encoders to low-delay mode.

3. Methodology

The top-level functional block diagram of the proposed training flow is shown in Fig. 1 and includes a preprocessor P, codec model M, target codec C and postprocessor O. Notably, the target codec is a traditional video codec, such as SVT-AV1 [32]. First, a source frame I_{in} is processed by P to generate a corresponding preprocessed frame I_P. The preprocessed frame is simultaneously encoded by M and C, which generate encoded frames I_M and I_C respectively. Finally, encoded frames are postprocessed with O to generate the output frame I_{out}. End-to-end distortion D is computed between I_{in} and I_{out} and the overall training loss is the weighted sum of D and the bitrate R_M. Notably, at inference, the codec model M is removed from the pipeline and only the target codec C is deployed for encoding.

3.1. Codec Model Pretraining

Given the target codec C in Fig. 1 is not differentiable, a mechanism for adapting the proposed pipeline to the target codec behavior is required. One option is black-box or derivative-free optimization methods [3, 7, 18]. However these methods typically rely on an approximation of the unknown function to be optimized, and are computationally infeasible for training at scale. Another option is to replicate the core prediction and transform blocks of the target codec in a differentiable manner [15, 27, 54]. Such approaches can model older codecs such as JPEG, or faster presets of AVC or HEVC successfully to an extent but fail to accurately model the operation of more sophisticated codecs such as AV1 or VVC.

Recent work [31, 57] has instead leveraged on neural image and video codecs for predicting the target codec behavior. In this work, a similar approach is followed, and a neural video codec M is adapted to model the rate-distortion behavior of a target codec C. Specifically, the scale-space flow architecture [2] is used to generate lossy frames, and the implicit encoder-decoder structure and associated loss functions are leveraged to align to the target codec behavior in a spatially localized manner. As per Agustsson *et al.* [2], for each of intra, flow and residual

encoding, the hyperprior based architecture of Balle *et al.* [5] is used, without any autoregressive components.

Notably, the hyperprior architecture comprises analysis and synthesis transforms, g_a and g_s, for encoding to and decoding from the quantized latent space $\tilde{y}$ respectively. A distortion loss $D(.; g_a, g_s)$ can thus be computed between source frame I_{in} (input to g_a) and reconstructed frame I_M (output by g_s). A hyper-analysis and hyper-synthesis transform (h_a and h_s respectively) constitute the hyperprior, which is used to model scale dependencies $\tilde{\sigma}$ within the latent space. The latent space $\tilde{y}$ is thus modelled as a Gaussian scale mixture on which the rate $R(.; h_a, h_s)$ can be optimized (as Shannon entropy). Traditionally, the architecture would then be trained with an aggregate loss function $L = D + \lambda R$, where λ represents a Lagrangian parameter for balancing between the rate and distortion components.

When performing neural codec alignment to a target codec, there are two key issues: *(i)* the model encoded frame must align to the target encoded frame (and not the source); *(ii)* selecting the best value for λ when trying to replicate target codec behavior is a non-trivial problem. Crucially, it is noted that the rate model is exclusively parameterized by h_a and h_s, and distortion exclusively by g_a and g_s. Subsequently, a novel two phase training methodology is proposed for aligning both rate and distortion to the target codec:

Phase 1: The first issue of target codec alignment is addressed via the distortion loss. To this end, the reference for the distortion loss is changed from the source frame I_{in} to the target codec frame I_C and the mean squared error is optimized:

$$D_M^{\text{hyp}}(I_M, I_C; \theta_g) = \mathbb{E}_{I_C}[||I_M - I_C||^2] \quad (1)$$

where θ_g represent the analysis and synthesis transform parameters.

Phase 2: Following alignment of g_a and g_s in Phase 1, the latent space $\tilde{y}$ is now conditioned on the target codec C. To circumvent selection of an optimal λ for rate, the implicit encoder-decoder structural prior of the aligned neural codec is leveraged in order to learn an unsupervised rate alignment to the target codec. Specifically, the maximum likelihood on the hyperprior model parameters is optimized, in order to learn the latent space distribution under target codec distortion alignment. The rate is thus optimized as:

$$R_M^{\text{hyp}}(I_M; \theta_h) = -\mathbb{E}_{\tilde{y},\tilde{z}}[\log p_{\tilde{y}|\tilde{z},I_C}(\tilde{y}|\tilde{z}, I_C) + \log p_{\tilde{z}|I_C}(\tilde{z}|I_C)] \quad (2)$$

where θ_h represents the hyperprior model parameters and $\tilde{z}$ is the quantized hyperprior latent space (which would also be encoded and transmitted as side information in a neural codec pipeline). For intra coding, the intra hyperprior model is optimized on intra encoded source frames using the two-phase strategy. For inter coding, the residual and flow hyperprior models are optimized jointly by summing the rate and distortion components in each phase.

The proposed two-phase pretraining ensures that R_M^{hyp} is close to the target codec R_C by performing an affine transform on R_M^{hyp}, i.e., $R_M = \alpha R_M^{\text{hyp}} + \beta$, where α and β are scalar learnable parameters optimized with MSE:

$$L_M^{(\alpha,\beta)}(R_M, R_C; \alpha, \beta) = \mathbb{E}_{R_C}[||R_M - R_C||^2]. \quad (3)$$

The similarity of R_M^{hyp} and R_C is required to properly balance the rate and distortion losses in end-to-end training. The aggregate rate and distortion for a sequence can thus be measured by computing the sum over all frames.

3.2. End-to-End Training

With the pretrained neural codec model, an end-to-end trainable pipeline with both the pre and postprocessor is developed, as illustrated in Fig. 1. In this case, the training loss is very similar to conventional neural codecs, but is now exclusively used to train the pre and postprocessers, i.e.:

$$L(I_{\text{out}}, I_{\text{in}}; \theta_P, \theta_O) = \mathbb{E}_{I_{\text{in}}}[D(I_{\text{in}}, I_{\text{out}}) + \lambda R_M] \quad (4)$$

where θ_P and θ_O are the pre and postprocessor parameters respectively, R_M is the codec model rate estimation (proportional to $R_M^{\text{flow}} + R_M^{\text{res}}$ for inter coding the flow and residual from motion compensation, and R_M^{intra} for intra coding) and D is the distortion loss component. Following subjective assessment tests, a weighted sum of: mean absolute error (MAE), SSIM, MS-SSIM and detail loss metric (DLM) * [37] is proposed for the distortion component D in Eq. (4). Specifically, $D(I_{\text{in}}, I_{\text{out}}) = \lambda_{\text{L1}}||I_{\text{in}} - I_{\text{out}}|| + \lambda_{\text{SSIM}}(1 - D_{\text{SSIM}}(I_{\text{in}}, I_{\text{out}})) + \lambda_{\text{MSSSIM}}(1 - D_{\text{MSSSIM}}(I_{\text{in}}, I_{\text{out}})) + \lambda_{\text{DLM}} D_{\text{DLM}}(I_{\text{in}}, I_{\text{out}})$.

Concept drift or domain shift for the codec model is an expected consequence of end-to-end training. The codec model is pretrained with natural videos whereas the preprocessor is continuously updated as training proceeds; thus, the data observed by the codec model has a non-stationary distribution. Since neural networks generally do not perform well on out-of-distribution data, an *iterative refinement* mechanism is proposed, where the codec model parameters θ_M are independently updated after each iteration of end-to-end training using the two-phase approach described previously. Notably, in this case the input to the target codec is now the output of the

*DLM is one of the core components of VMAF [41] and is found to better preserve details under visual degradation. Given that it is primarily wavelet based, it can be implemented in a differentiable manner and used for model optimization.

preprocessor I_P and not the source frame, and I_C and R_C represent the frame and rate output by the target codec.

Additionally, the gradients backpropagated to the preprocessor can be better aligned towards the target codec via the *stop-gradient* operator. First, it is assumed for the remaining discussion in the section that the tensors representing the frames are flattened, i.e., $\{I_P, I_M, I_C\} \in \mathbb{R}^{N\times 1}$, where N is the number of pixels in the tensor. If the target codec C was differentiable then the accuracy of the gradients backpropagated will increase by decreasing:

$$G = |J_M(I_P)\nabla_{I_M} L - J_C(I_P)\nabla_{I_C} L|, \tag{5}$$

where $J_M(I_P) \in \mathbb{R}^{N\times N}$ is the transposed Jacobian of M w.r.t. I_P and $\nabla_{I_M} L$ is the gradient of the objective w.r.t. I_M. Assuming that the frames I_C generated by the target codec C are input to the postprocessor O then Eq. (5) becomes

$$G = |(J_M(I_P) - J_C(I_P))\nabla_{I_C} L|, \tag{6}$$

Since $\nabla^{(j)}_{I_C} L$ is necessary for the networks to train, $J^{(i,j)}_M(I_P) - J^{(i,j)}_C(I_P)$ should be close to zero. Thus,

$$\forall i, j \in [1, N], \left|\frac{\partial I^{(i)}_M}{\partial I^{(j)}_P} - \frac{\partial I^{(i)}_C}{\partial I^{(j)}_P}\right| = \left|\frac{\partial}{\partial I^{(j)}_P}(I^{(i)}_M - I^{(i)}_C)\right| \tag{7}$$

should be minimized. The above analysis can be extended to the rate component. Eq. (7) shows that accurate gradients can be backpropagated to the preprocessor if the codec model's output is close to the target codec's output and the input to the postprocessor is from the target video codec. In the proposed training regime, the input to the postprocessor is therefore set as

$$I_{O_{\text{in}}} = \overline{(I_C - I_M)} + I_M, \tag{8}$$

where $\bar{a}$ indicates the stop-gradient operator [63] applied to the tensor a and I_C is the output of the non-differentiable video codec. Eq. (8) allows frames generated by the target codec to be passed to the postprocessor while using the codec model for backpropagation.

In summary, the proposed training procedure has two main steps; iterative update of the codec model and the end-to-end update of the pre- and postprocessors. For the end-to-end update, a stop-gradient-based approach is proposed. Simple CNNs are used for pre and postprocessing. Details of the pre and postprocessor architecture are in the supplementary. The end-to-end training pipeline is shown in Algorithm 1.

Algorithm 1 End-to-end training pipeline of the pre and postprocessor

Require: $\boldsymbol{I}_{\text{in}}$, batch of video frames
Require: θ_M, parameters of the codec model
Require: θ_P, parameters of the preprocessor
Require: θ_O, parameters of the postprocessor
Require: α_L, learning rate for training
for all $\boldsymbol{I}_{\text{in}}$ **do**
  $\boldsymbol{I}_P \leftarrow P(\boldsymbol{I}_{\text{in}})$
  $(\boldsymbol{I}_C, \boldsymbol{R}_C) \leftarrow C(\boldsymbol{I}_P)$
  $(\boldsymbol{I}_M, \boldsymbol{R}_M) \leftarrow M(\boldsymbol{I}_P)$
  Get D^{hyp}_M and $\theta_g \leftarrow \theta_g - \alpha_L \nabla_{\theta_g} D^{hyp}_M$
  Get R^{hyp}_M and $\theta_h \leftarrow \theta_h - \alpha_L \nabla_{\theta_h} R^{hyp}_M$
  Get $L^{(\alpha,\beta)}_M$ and $(\alpha, \beta) \leftarrow (\alpha, \beta) - \alpha_L \nabla_{(\alpha,\beta)} L^{(\alpha,\beta)}_M$
  Generate $\boldsymbol{I}_{O_{\text{in}}}$ using Eq. (8)
  $\boldsymbol{I}_{\text{out}} \leftarrow O(\boldsymbol{I}_{O_{\text{in}}})$
  Get L using Eq. (4)
  $\theta_P \leftarrow \theta_P - \alpha_L \nabla_{\theta_P} L$
  $\theta_O \leftarrow \theta_O - \alpha_L \nabla_{\theta_O} L$
end for

4. Experimental Results

4.1. Training

The Vimeo-90k dataset [70] was used to train the codec model and perform end-to-end training. The training pipeline and the pre- and postprocessor are implemented in the YUV colorspace. A fixed learning rate of 10^{-4} is used for each network with the rest of hyper-parameters set as: $\lambda = 0.1$; $\lambda_{\text{L1}} = 0.1$, $\lambda_{\text{SSIM}} = 0.25$; $\lambda_{\text{MSSSIM}} = 0.25$; $\lambda_{\text{DLM}} = 0.2$. Batch size, GOP and maximum iterations are set to 4, 4 and 500000, respectively. Both Phase 1 and 2 of pretraining are performed for 500000 iterations. The model is trained on source frames only in an open-loop configuration; i.e., no reconstructed frames are fed into the preprocessor as input. A low-delay recipe of SVT-AV1 (described in supplementary) is used for codec model alignment for pre-training and end-to-end training. The codec model encoding is aligned to the GOP of the target codec, such that intra and inter coding are also aligned. For experiments on VVC, it was found that codec model alignment to a low-delay recipe of VVenC provided slight BD-rate gains over SVT-AV1 alignment, albeit at high computational overhead during training. Therefore, the presented results are carried out with a single neural wrapper model (one encode) developed based on codec proxy model alignment to SVT-AV1. This is practically more feasible than [24], which requires 9 combinations of models (9 encodes) + convex hull optimization and is thus far more computationally demanding.

4.2. Evaluation Protocol

Four 1080p video datasets were used for evaluation: *(i)* 16 gaming sequences captured in a commercial setup (**Gaming**); *(ii)* The 16 XIPH natural sequences used by recent work [11] (**XIPH**); *(iii)* The 1080p sequences of the UVG [48] dataset (**UVG**); *(iv)* HEVC-B [8] dataset

(**HEVC-B**). The proposed method was applied with SVT-AV1 [32] for AV1 and VVenC [10] for VVC. All experiments were performed with YUV 4:2:0 8-bit sequences. In all cases, the protocol of Li *et al.* [35] is followed and 96 frames are evaluated. However, unlike Li [35], both standard encoders are used *without* their low-delay limitation, in order to attain their best performance (see supplementary for details on sequences and encoding recipes). Ablation results are with the SVT-AV1 lowdelay recipe (see supplementary). Results in Tab. 3 and 4 are on the Gaming dataset.

In order to cover a wide range of well-established objective quality scoring methods, the following scores are used: SSIM, MS-SSIM, VMAF [42, 52] and the AVQT score from Apple [4]. In order to showcase the generality of the proposed approach, the same pre- and postprocessor models are used for: *(i)* all the points on the RD curves, *(ii)* all quality scores and *(iii)* both standard codecs. Perceptual models are used for competing methods when available. Comparison is provided against a competitive video preprocessor (DPP [11]), SOTA neural codecs (DCVC-DC [35] and DCVC-FM [36]), and perceptual modes of standard codecs. The comparisons include results of end-to-end training with the codec model of Hu *et al.* [24] with the proposed network architectures and losses used to illustrate the effectiveness of the proposed codec model. For DCVC-DC and DCVC-FM, the best available pretrained models (publicly provided by the authors) are used.

4.3. Objective Quality Assessment

The quantitative results of the proposed method are shown in[†] Fig. 2, and Tab. 1 and Tab. 2. These results show that the proposed approach provides for BD-rates between -5.4% to -30% in comparison to the underlying standard encoders. BD-rate savings tend to be higher for quality scores that are more perceptually-oriented. Compared to competing methods, the proposed approach is shown to offer significantly more consistent BD-rate improvements over *all* quality scoring methods (ALL columns of Tab. 1 and Tab. 2). Importantly, as shown in Fig. 2, unlike DCVC-DC that is only covering a very limited bitrate-quality range at the very top end, the proposed approach outperforms all other methods on the entire "active" region of bitrate-quality, i.e., VMAF of 40-93 [32], which encapsulates the entirety of commercially-relevant bitrates for 1080p video streaming. Given that the neural wrapper is optimizing perceptually, BD-rates for PSNR tend to show regression on PSNR (see supplementary) but visual inspection and an array of perceptual metrics indicate that the proposed approach offers significant benefits

[†]All combined plots of Fig. 2 and Fig. 3 are produced with the slope-based integration method of Wu *et al.* [68] and are shown in log-scale.

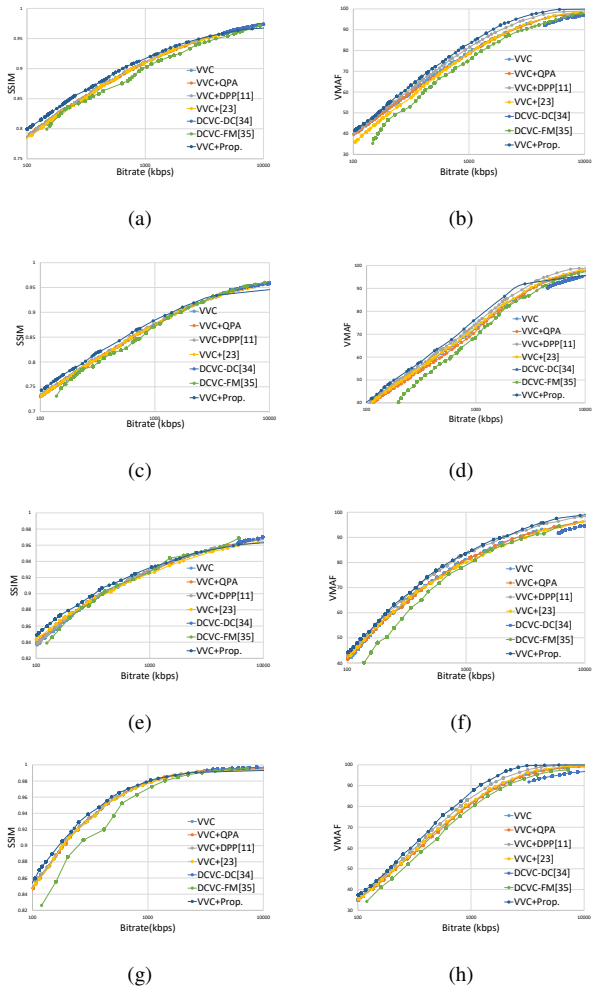


Figure 2. Combined rate-quality plots over Gaming ((a), (b)), XIPH ((c), (d)), UVG ((e), (f)) and HEVC-B ((g), (h)) datasets. See supplementary for more results.

in visual quality. Nonetheless, the results of DCVC-DC and DCVC-FM are remarkably good given that they are fully neural approaches.

4.4. Subjective Quality Assessment

To assess the visual quality of the proposed approach, a 5-scale absolute categorical rating (ACR) test with hidden reference was conducted under the ITU-T P.910 test conditions and SUREAL post-processing [40, 43] (see supplementary for details). Raters were shown videos generated by the codec and proposed method. Separate tests were conducted for AV1 and VVC with each test comprising 72 sequences. The combined rate-MOS plots are shown in Fig. 3. The BD-rates reported in the figure for the proposed approach are mostly in line with AVQT and VMAF BD-rates reported for Gaming in Tab. 1 and

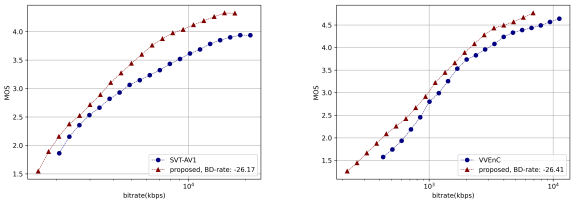
Table 1. BD-rates vs. AV1 (more negative is better). AV1+SSIM denotes codec tuned for SSIM (-tune ssim). ALL indicates average BD-rate over all quality scores.

Dataset	Method	SSIM	MSSSIM	AVQT	VMAF	ALL
Gaming	AV1+SSIM	-4.5	-5.4	-5.1	2.1	-3.2
	AV1+DPP [11]	-5.2	-6.1	-17.6	-11.3	-10.0
	AV1+[24]	14.9	6.3	12.2	-1.4	8.0
	DCVC-DC [35]	-15.4	**-36.1**	**-35.6**	21.84	-16.3
	DCVC-FM [36]	14.8	6.0	4.2	-2.9	5.5
	AV1+Proposed	**-15.5**	-13.8	-19.6	**-30.0**	**-19.7**
XIPH	AV1+SSIM	-3.5	-4.0	-5.4	3.2	-2.4
	AV1+DPP [11]	-4.3	-3.9	-9.0	-16.7	-8.5
	AV1+[24]	12.9	6.1	4.5	1.7	6.3
	DCVC-DC [35]	**-18.4**	**-25.5**	12.0	32.2	0.3
	DCVC-FM [36]	17.6	9.0	8.2	3.5	9.6
	AV1+Proposed	-10.9	-7.1	**-28.8**	**-30.1**	**-19.2**
UVG	AV1+SSIM	-2.8	-3.1	-2.8	2.9	-1.4
	AV1+DPP [11]	-3.8	-3.0	-9.2	-21.7	-9.4
	AV1+[24]	9.1	4.4	3.3	3.2	5.0
	DCVC-DC [35]	**-20.8**	**-10.2**	4.2	18.2	-2.1
	DCVC-FM [36]	20.4	9.6	9.8	9.4	12.3
	AV1+Proposed	-16.2	-9.7	**-22.2**	**-31.9**	**-20.0**
HEVC-B	AV1+SSIM	-5.4	-5.3	-4.7	2.9	-3.1
	AV1+DPP [11]	-8.7	-6.9	-5.7	-20.1	-10.3
	AV1+[24]	2.6	1.8	-3.1	1.1	0.6
	DCVC-DC [35]	**-35.3**	**-29.1**	-11.6	37.2	-9.7
	DCVC-FM [36]	-6.5	-11.7	-7.7	-5.8	-7.9
	AV1+Proposed	-14.5	-13.4	**-15.1**	**-41.8**	**-21.2**

Table 2. BD-rates vs. VVC (more negative is better). VVC+QPA indicates quantization parameter adaptation is enabled. ALL indicates average BD-rate over all quality scores.

Dataset	Method	SSIM	MSSSIM	AVQT	VMAF	ALL
Gaming	VVC+QPA	0.4	4.6	13.4	1.5	4.9
	VVC+DPP [11]	-1.3	-1.6	-9.2	-9.0	-5.3
	VVC+[24]	39.5	32.3	11.4	7.8	22.7
	DCVC-DC [35]	9.1	**-11.5**	**-21.3**	38.9	3.8
	DCVC-FM [36]	4.9	18.4	10.9	18.4	13.1
	VVC+Proposed	**-16.6**	-8.7	-19.0	**-14.3**	**-14.6**
XIPH	VVC+QPA	-0.9	2.2	1.1	2.7	1.3
	VVC+DPP [11]	-2.3	-1.6	-10.5	-10.4	-6.2
	VVC+[24]	29.4	24.4	12.2	9.5	18.9
	DCVC-DC [35]	-8.1	**-12.0**	9.2	41.1	7.5
	DCVC-FM [36]	4.3	13.1	0.5	16.8	8.7
	VVC+Proposed	**-11.4**	-5.4	**-16.5**	**-18.0**	**-12.8**
UVG	VVC+QPA	-0.1	1.6	7.7	1.5	2.7
	VVC+DPP [11]	-1.4	-1.1	-7.0	-11.4	-5.2
	VVC+[24]	26.8	23.5	7.9	6.5	16.2
	DCVC-DC [35]	-8.2	16.0	-1.4	29.1	8.9
	DCVC-FM [36]	-2.7	14.1	12.5	14.7	9.65
	VVC+Proposed	**-11.8**	**-5.4**	**-13.9**	**-13.8**	**-11.2**
HEVC-B	VVC+QPA	0.9	1.2	1.1	2.1	1.3
	VVC+DPP [11]	-19.2	-16.7	-17.2	-16.7	-17.4
	VVC+[24]	0.5	0.5	0.1	-2.3	-0.3
	DCVC-DC [35]	-17.0	-5.5	6.3	49.9	8.4
	DCVC-FM [36]	18.8	15.8	14.4	12.1	15.3
	VVC+Proposed	**-31.8**	**-27.4**	**-25.2**	**-33.3**	**-29.4**

Tab. 2. Sample visual results of the proposed method are shown in Fig. 4. The results show that the proposed approach is temporally coherent, and better capable of preserving structure, retaining textures and denoising codec artifacts, also reflected in the MOS scores of Fig. 3.



(a) (b)

Figure 3. Combined rate-quality plots with P.910 MOS.

Table 3. BD-rates of different component configurations.

Pre	Post	SSIM	MSSSIM	AVQT	VMAF
✓	✓	-25.6	-20.8	-21.4	-20.5
✓	✗	17.9	13.2	29.8	-34.1
✗	✓	-2.9	0.4	4.1	-33.3

4.5. Ablation

Two-phase pretraining: The rate-quality alignment between target codec C and model M after two-phase pretraining is validated by comparing their performance on a dataset ov dieo sequences. The Pearson correlation coefficient (PCC) is improved from 0.56 and 0.60 to 0.88 and 0.85 for bitrate and PSNR, respectively (see supplementary for more results). Note that the PCC for rate is computed between R_C and R_M^{hyp}. This indicates that both distortion and bitrate of the codec model and target codec correlate better after the two-phase pretraining, which is a sufficient condition for gradient-based optimization [31]. Tab. 4 shows the effect of pretraining on wrappers.

Pre/post processing: Tab. 3 shows the effect of removing either the pre or the postprocessor from the overall pipeline under the same training pipeline. Removing the preprocessor is equivalent to training a video denoiser. While BD-rate gains for VMAF are maintained with only the pre- and postprocessor, this can be attributed to sharpening effects over the source content, which VMAF has susceptibility to [1]. Indeed, with more naive content enhancement such as sharpening, fidelity oriented methods such as SSIM and MS-SSIM will experience BD-rate losses, which can be seen when training both pre and postprocessors alone. Only jointly training both pre- and postprocessors is able to achieve substantial BD-rate gains across all quality scores. Fig. 5 provides a generic example where the preprocessor embeds information into the source frame which does not significantly increase the bitrate and can survive the codec compression. This information is used by the postprocessor for reconstruction.

End-to-end training: For end-to-end training, an ablation is performed over the proposed iterative updates of the codec model and stop-gradient during training. The results are provided in Tab. 4 under different training con-

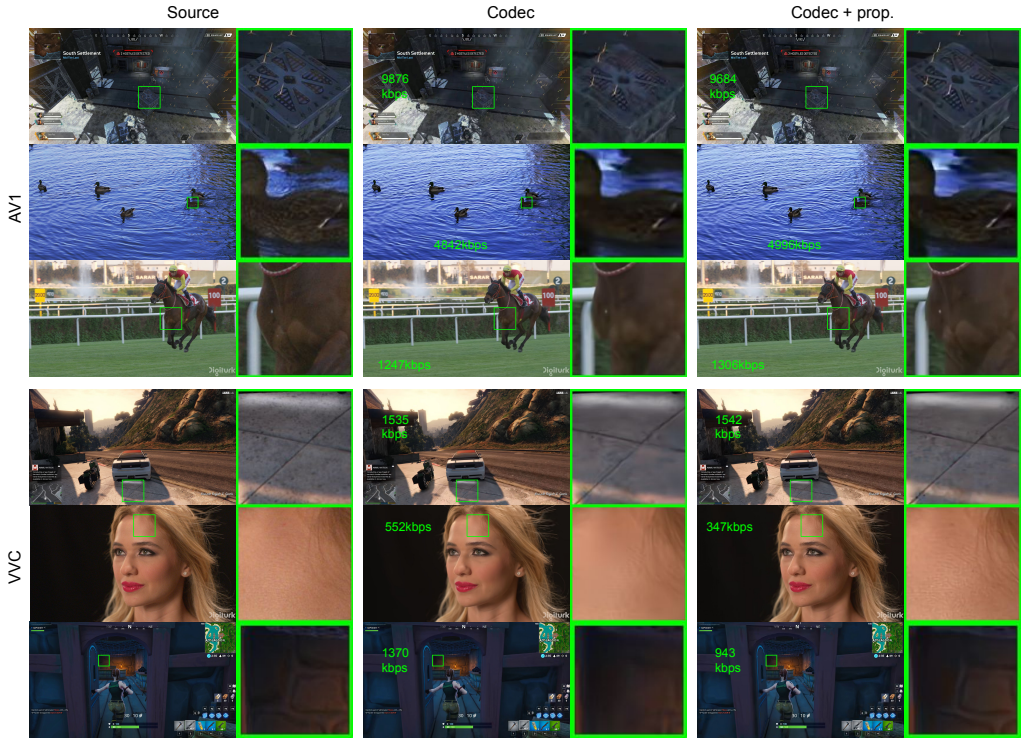


Figure 4. Qualitative comparison of codecs with and without the proposed method. Zoom in for better viewing.

Figure 5. Results of components of the pipeline where (a), (b), (c) and (d) are the source frame, preprocessor output, codec output and postprocessor output, respectively.

Table 4. BD-rates with different training configurations.

Pretrain	Iter. Update	Stop Grad.	SSIM	MSSSIM	AVQT	VMAF
✓	✓	✓	-25.6	-20.8	-21.4	-20.5
✗	✓	✓	-23.1	-14.4	-17.3	-18.7
✓	✓	✗	16.4	0.01	10.5	-15.3
✓	✗	✓	-11.2	-6.5	-7.7	-13.1

figurations. Notably, when training the pre- and postprocessor with only iterative updates, there are BD-rate losses in SSIM and MS-SSIM. This is a by-product of the domain shift between the source and preprocessor output that increases as end-to-end training progresses. When performing end-to-end training with both iterative updates and stop gradient, there is a sufficient alignment constraint on the codec model, which leads to demonstrable BD-rate gains across quality scoring methods.

5. Conclusion

This work proposes a new approach to jointly train neural pre and postprocessors to improve the rate-quality performance of conventional video codecs. It is proposed to use a neural codec to model the standard codec's sophisticated rate-quality characteristics, combined with: *(i)* a novel approach to pretrain the codec model; *(ii)* a new method to train the neural wrapper model jointly with the codec model in the middle. Extensive evaluations with state-of-the-art AV1 and VVC codecs show that the proposed neural wrapper approach provides for consistent BD-rate improvement over standard codecs. Unlike neural methods for preprocessing and encoding, the proposed method allows for *consistent* improvement over all quality scores (SSIM, MSSSIM, AVQT, VMAF, P.910 MOS) over the *entire* rate-quality region of modern standards.

References

[1] On vmaf's property in the presence of image enhancement operations. https://docs.google.com/document/d/1dJczEhXO0MZjBSNyKmd3ARiCTdFVMNPBykH4_HMPoyY/edit#heading=h.oaikhnw46pw5. Last accessed: 2020-11-16. 7

[2] Eirikur Agustsson, David Minnen, Nick Johnston, Johannes Balle, Sung Jin Hwang, and George Toderici. Scale-space flow for end-to-end optimized video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2020. 2, 3

[3] Stéphane Alarie, Charles Audet, Aïmen E Gheribi, Michael Kokkolaras, and Sébastien Le Digabel. Two decades of blackbox optimization applications. *EURO Journal on Computational Optimization*, 9:100011, 2021. 3

[4] Apple. Apple avqt presentation", 2021. https://developer.apple.com/videos/play/wwdc2021/10145. 1, 6

[5] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018. 4

[6] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 1

[7] Mohammad Reza Bonyadi and Zbigniew Michalewicz. Particle swarm optimization for single objective continuous space problems: a review. *Evolutionary computation*, 25(1):1–54, 2017. 3

[8] Frank Bossen et al. Common test conditions and software reference configurations. *JCTVC-L1100*, 12(7):1, 2013. 5

[9] Eirina Bourtsoulatze, Aaron Chadha, Ilya Fadeev, Vasileios Giotsas, and Yiannis Andreopoulos. Deep video precoding. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 2

[10] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021. 6

[11] Aaron Chadha and Yiannis Andreopoulos. Deep perceptual preprocessing for video coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14852–14861, 2021. 2, 5, 6, 7

[12] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4947–4956, 2021. 3

[13] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022. 3

[14] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5962–5971, 2022. 3

[15] Jinyoung Choi and Bohyung Han. Task-aware quantization network for jpeg image compression. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 309–324. Springer, 2020. 3

[16] Jianing Deng, Li Wang, Shiliang Pu, and Cheng Zhuo. Spatio-temporal deformable convolution for compressed video quality enhancement. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10696–10703, 2020. 3

[17] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3476–3485. IEEE, 2019. 3

[18] Ryosuke Furuta, Naoto Inoue, and Toshihiko Yamasaki. Pixelrl: Fully convolutional network with reinforcement learning for image processing. *IEEE Transactions on Multimedia*, 22(7):1704–1719, 2019. 3

[19] Vasileios Giotsas, Nikos Deligiannis, Pam Fisher, and Yiannis Andreopoulos. Perceptual video quality estimation by regression with myopic experts. In *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2015. 1

[20] Onur G Guleryuz, Philip A Chou, Hugues Hoppe, Danhang Tang, Ruofei Du, Philip Davidson, and Sean Fanello. Sandwiched image compression: wrapping neural networks around a standard codec. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3757–3761. IEEE, 2021. 3

[21] Onur G Guleryuz, Philip A Chou, Hugues Hoppe, Danhang Tang, Ruofei Du, Philip Davidson, and Sean Fanello. Sandwiched image compression: Increasing the resolution and dynamic range of standard codecs. In *2022 Picture Coding Symposium (PCS)*, pages 175–179. IEEE, 2022. 3

[22] Mengxi Guo, Shijie Zhao, Hao Jiang, Junlin Li, and Li Zhang. Video compression with arbitrary rescaling network. *arXiv preprint arXiv:2306.04202*, 2023. 3

[23] Yung-Han Ho, Chih-Peng Chang, Peng-Yu Chen, Alessandro Gnutti, and Wen-Hsiao Peng. Canf-vc: Conditional augmented normalizing flows for video compression. In *European Conference on Computer Vision*, pages 207–223. Springer, 2022. 2

[24] Yueyu Hu, Chenhao Zhang, Onur G Guleryuz, Debargha Mukherjee, and Yao Wang. Standard compliant video coding using low complexity, switchable neural wrappers. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 1922–1928. IEEE, 2024. 3, 5, 6, 7

[25] Zhihao Hu and Dong Xu. Complexity-guided slimmable decoder for efficient deep video compression. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14358–14367, 2023. 2

[26] Quan Huynh-Thu and Mohammed Ghanbari. The accuracy of psnr in predicting video quality for different video scenes and frame rates. *Telecommunication Systems*, 49: 35–48, 2012. 1

[27] Berivan Isik, Onur G Guleryuz, Danhang Tang, Jonathan Taylor, and Philip A Chou. Sandwiched video compression: Efficiently extending the reach of standard codecs with neural wrappers. *arXiv preprint arXiv:2303.11473*, 2023. 3

[28] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 645–660. Springer, 2020. 3

[29] Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Ya-Li Li, Shengjin Wang, and Qi Tian. Video super-resolution with temporal group attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8008–8017, 2020. 3

[30] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3224–3232, 2018. 3

[31] Jan P Klopp, Keng-Chi Liu, Liang-Gee Chen, and Shao-Yi Chien. How to exploit the transferability of learned image compression to conventional codecs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16165–16174, 2021. 2, 3, 7

[32] Faouzi Kossentini, Hassen Guermazi, Nader Mahdi, Chekib Nouira, Amir Naghdinezhad, Hassene Tmar, Omar Khlif, Phoenix Worth, and Foued Ben Amara. The svt-av1 encoder: overview, features and speed-quality tradeoffs. In *Applications of Digital Image Processing XLIII*, page 1151021. International Society for Optics and Photonics, 2020. 3, 6

[33] Théo Ladune, Pierrick Philippe, Wassim Hamidouche, Lu Zhang, and Olivier Déforges. Conditional coding for flexible learned video compression. *arXiv preprint arXiv:2104.07930*, 2021. 2

[34] Jiahao Li, Bin Li, and Yan Lu. Hybrid spatial-temporal entropy modelling for neural video compression. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1503–1511, 2022. 2

[35] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with diverse contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22616–22626, 2023. 2, 6, 7

[36] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with feature modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26099–26108, 2024. 2, 6, 7

[37] Songnan Li, Fan Zhang, Lin Ma, and King Ngi Ngan. Image quality assessment by separately evaluating detail losses and additive impairments. *IEEE Transactions on Multimedia*, 13(5):935–949, 2011. 4

[38] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 335–351. Springer, 2020. 3

[39] Yinxiao Li, Pengchong Jin, Feng Yang, Ce Liu, Ming-Hsuan Yang, and Peyman Milanfar. Comisr: Compression-informed video super-resolution. in 2021 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 2523–2532, 2021. 3

[40] Zhi Li and Christos G Bampis. Recover subjective quality scores from noisy measurements. In *2017 Data compression conference (DCC)*, pages 52–61. IEEE, 2017. 6

[41] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 6, 2016. 4

[42] Zhi Li, Christos Bampis, Julie Novak, Anne Aaron, Kyle Swanson, Anush Moorthy, and J Cock. Vmaf: The journey continues. *Netflix Technology Blog*, 2018. 1, 6

[43] Zhi Li, Christos G Bampis, Lukáš Krasula, Lucjan Janowski, and Ioannis Katsavounidis. A simple model for subject behavior in subjective experiments. *arXiv preprint arXiv:2004.02067*, 2020. 6

[44] Jianping Lin, Dong Liu, Houqiang Li, and Feng Wu. M-lvc: Multiple frames prediction for learned video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3546–3554, 2020. 2

[45] Jerry Liu, Shenlong Wang, Wei-Chiu Ma, Meet Shah, Rui Hu, Pranaab Dhawan, and Raquel Urtasun. Conditional entropy coding for efficient video compression. In *European Conference on Computer Vision*, pages 453–468. Springer, 2020. 2

[46] 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 Conference on Computer Vision and Pattern Recognition*, pages 11006–11015, 2019. 2

[47] Fabian Mentzer, George Toderici, David Minnen, Sung-Jin Hwang, Sergi Caelles, Mario Lucic, and Eirikur Agustsson. Vct: A video compression transformer. *arXiv preprint arXiv:2206.07307*, 2022. 2

[48] Alexandre Mercat, Marko Viitanen, and Jarno Vanne. Uvg dataset: 50/120fps 4k sequences for video codec analysis and development. In *Proceedings of the 11th ACM Multimedia Systems Conference*, pages 297–302, 2020. 5

[49] Reza Pourreza, Hoang Le, Amir Said, Guillaume Sautiere, and Auke Wiggers. Boosting neural video codecs by exploiting hierarchical redundancy. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5355–5364, 2023. 2

[50] Zhongwei Qiu, Huan Yang, Jianlong Fu, and Dongmei Fu. Learning spatiotemporal frequency-transformer for compressed video super-resolution. In *European Conference on Computer Vision*, pages 257–273. Springer, 2022. 3

[51] Alexander Raake, Silvio Borer, Shahid M Satti, Jörgen Gustafsson, Rakesh Rao Ramachandra Rao, Stefano Medagli, Peter List, Steve Göring, David Lindero, Werner Robitza, et al. Multi-model standard for bitstream-, pixel-based and hybrid video quality assessment of uhd/4k: Itu-t p. 1204. *IEEE Access*, 8:193020–193049, 2020. 1

[52] Reza Rassool. Vmaf reproducibility: Validating a perceptual practical video quality metric. In *2017 IEEE international symposium on broadband multimedia systems and broadcasting (BMSB)*, pages 1–2. IEEE, 2017. 6

[53] Oren Rippel, Alexander G Anderson, Kedar Tatwawadi, Sanjay Nair, Craig Lytle, and Lubomir Bourdev. Elf-vc: Efficient learned flexible-rate video coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14479–14488, 2021. 2

[54] Amir Said, Manish Kumar Singh, and Reza Pourreza. Differentiable bit-rate estimation for neural-based video codec enhancement. In *2022 Picture Coding Symposium (PCS)*, pages 379–383. IEEE, 2022. 3

[55] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6626–6634, 2018. 3

[56] Xihua Sheng, Jiahao Li, Bin Li, Li Li, Dong Liu, and Yan Lu. Temporal context mining for learned video compression. *IEEE Transactions on Multimedia*, 2022. 2

[57] Hanbin Son, Taeoh Kim, Hyeongmin Lee, and Sangyoun Lee. Enhanced standard compatible image compression framework based on auxiliary codec networks. *IEEE Transactions on Image Processing*, 31:664–677, 2021. 3

[58] Yannick Strümpler, Ren Yang, and Radu Timofte. Learning to improve image compression without changing the standard decoder. In *European Conference on Computer Vision*, pages 200–216. Springer, 2020. 2

[59] Hossein Talebi and Peyman Milanfar. Learned perceptual image enhancement. In *2018 IEEE International Conference on Computational Photography (ICCP)*, pages 1–13. IEEE, 2018. 2

[60] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 4472–4480, 2017. 3

[61] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu Tdan. temporally-deformable alignment network for video super-resolution. in 2020 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3357–3366, 2020. 3

[62] Yuan Tian, Guo Lu, Xiongkuo Min, Zhaohui Che, Guangtao Zhai, Guodong Guo, and Zhiyong Gao. Self-conditioned probabilistic learning of video rescaling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4490–4499, 2021. 3

[63] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 5

[64] Ties van Rozendaal, Tushar Singhal, Hoang Le, Guillaume Sautiere, Amir Said, Krishna Buska, Anjuman Raha, Dimitris Kalatzis, Hitarth Mehta, Frank Mayer, et al. Mobilenvc: Real-time 1080p neural video compression on a mobile device. *arXiv preprint arXiv:2310.01258*, 2023. 2

[65] Abhinau K Venkataramanan, Chengyang Wu, Alan C Bovik, Ioannis Katsavounidis, and Zafar Shahid. A hitchhiker's guide to structural similarity. *IEEE Access*, 9: 28872–28896, 2021. 1

[66] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 3

[67] Yingwei Wang, Takashi Isobe, Xu Jia, Xin Tao, Huchuan Lu, and Yu-Wing Tai. Compression-aware video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2012–2021, 2023. 3

[68] Ping-Hao Wu, Ioannis Katsavounidis, Zhijun Lei, David Ronca, Hassene Tmar, Omran Abdelkafi, Colton Cheung, Foued Ben Amara, and Faouzi Kossentini. Towards much better svt-av1 quality-cycles tradeoffs for vod applications. In *Applications of Digital Image Processing XLIV*, pages 236–256. SPIE, 2021. 6

[69] Li Xu, Gang He, Jinjia Zhou, Jie Lei, Weiying Xie, Yunsong Li, and Yu-Wing Tai. Transcoded video restoration by temporal spatial auxiliary network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2875–2883, 2022. 3

[70] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 3, 5

[71] Ren Yang, Mai Xu, and Zulin Wang. Decoder-side hevc quality enhancement with scalable convolutional neural network. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 817–822. IEEE, 2017. 2

[72] Ren Yang, Mai Xu, Tie Liu, Zulin Wang, and Zhenyu Guan. Enhancing quality for hevc compressed videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(7):2039–2054, 2018. 2

[73] Ren Yang, Mai Xu, Zulin Wang, and Tianyi Li. Multi-frame quality enhancement for compressed video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6664–6673, 2018. 2

[74] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3106–3115, 2019. 3

[75] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, Tao Lu, Xin Tian, and Jiayi Ma. Omniscient video super-resolution. In *Proceedings of the IEEE/CVF International*

Conference on Computer Vision, pages 4429–4438, 2021. 3

[76] Hengsheng Zhang, Xueyi Zou, Jiaming Guo, Youliang Yan, Rong Xie, and Li Song. A codec information assisted framework for efficient compressed video super-resolution. In *European Conference on Computer Vision*, pages 220–235. Springer, 2022. 3