INSTANCE-ADAPTIVE VIDEO COMPRESSION: IMPROVING NEURAL CODECS BY TRAINING ON THE TEST SET

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a video compression algorithm based on instance-adaptive learning. On each video sequence to be transmitted, we finetune a pretrained compression model. The optimal parameters are transmitted to the receiver along with the latent code. By entropy-coding the parameter updates under a suitable mixture model prior, we ensure that the network parameters can be encoded efficiently. This instance-adaptive compression algorithm is agnostic about the choice of base model and has the potential to improve any neural video codec. On UVG, HEVC, and Xiph datasets, our codec improves the performance of a low-latency scale-space flow model by between 24 % and 26 % BD-rate savings, and that of a state-of-the-art B-frame model by 17 to 20 % BD-rate savings. We also demonstrate that instance-adaptive finetuning improves the robustness to domain shift. Finally, our approach reduces the capacity requirements on compression models. We show that it enables a state-of-the-art performance even after reducing the network size by 72 %.

1 INTRODUCTION

Neural compression methods have enormous potential to improve the efficiency of video coding. With video constituting the large majority of internet traffic, this has significant implications for the internet at large (Sandvine, 2019). State-of-the-art algorithms model each frame by warping the previous one with a neural estimate of the optical flow (Lu et al., 2019) or scale-space flow (Agustsson et al., 2020) and adding residuals modeled by another network. Both the optical flow and the residuals are compressed with variational autoencoders. Such neural codecs have recently achieved results on par with popular classical codecs (Agustsson et al., 2020; Rippel et al., 2021; Pourreza & Cohen, 2021) such as H.264 / AVC (Wiegand et al., 2003) and H.265 / HEVC (Sullivan et al., 2012). However, relatively little research has focused on their computational complexity, and matching the rate-distortion performance of H.266 / VVC (Bross et al., 2018) is still an open problem.

Neural video codecs depend critically on *generalization*: they are developed assuming that a good performance on training and validation datasets translates to a good performance at test time. However, this assumption does not always hold in practice, for instance because of limited training data or imperfect optimization. Domain shift is a particularly challenging problem given the variety in video content and styles; for instance, neural video codecs trained on natural scene video data often perform poorly on animated sequences (Agustsson et al., 2020).

Here we address this problem by optimizing the autoencoder — finetuning its weights — for each single video to be compressed at test time. After training on the test set in this way, the finetuned network parameters need to be made available on the receiver side: they are compressed, quantized, and entropy-coded under a suitable prior to the bitstream along with the latent code. Optimizing the compression network for every video in this way relaxes the requirement of training data that closely matches the test distribution as well as the necessary model capacity. It also allows trading off encoding compute for better compression efficiency.

The essential idea of instance-adaptive finetuning was recently proposed by Rozendaal et al. (2021). The authors demonstrated the idea on I-frame compression, i. e. compressing a set of similar images, but did not apply it to video sequences yet. In this work we extend the idea to the compression of full

videos, in which only periodic key frames are compressed as images and most frames are encoded relative to one or two reference frames. Our method is general and can be applied to various settings and base models. We first demonstrate it in a low-latency setting with I-frames and P-frames using a scale-space flow architecture (Agustsson et al., 2020) as base model. Next we show its performance in a B-frame setting, using the base model proposed by Pourreza & Cohen (2021).

On the UVG-1k, HEVC class-B, and Xiph-5N datasets, our new instance-adaptive video compression codec yields BD-rate savings of 17 to 26 % over the respective base models and 6 to 44 % over the popular ffmpeg (x265) implementation of the H.265 codec (FFmpeg; VideoLAN, b). In addition, instance-adaptive finetuning lends itself to a possible reduction in model size, because a smaller network may suffice to model a single instance. We show that in this framework smaller models can still outperform current state-of-the-art neural codecs while reducing the computational complexity of the decoder by 72%. Overall, we demonstrate that unlike other neural video codecs, our method can trade off encoding compute against compression performance, and that this trade-off is more effective than for classical codecs.

2 Related work

Neural video compression The standard framework used by most neural compression codecs (either implicitly or explicitly) is that of the variational (Kingma & Welling, 2014) or compressive (Theis et al., 2017) autoencoder. An encoder (or approximate posterior) $q_\phi(z|x)$ maps a data point x to a latent representation z. This latent is transmitted to the decoding party by means of entropy coding under a latent prior $p_\theta(z)$. The receiver can then reconstruct the sample with a decoder (or likelihood) $p_\theta(x|z)$. The encoder, prior, and decoder are neural networks, parameterized by weights ϕ, θ as indicated by the subscripts.

These models are trained by minimizing the rate-distortion loss

$$\mathcal{L}_{\text{RD}}(\phi,\theta) = \mathbb{E}_{x\sim p(x)}\left[\beta \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}\left[-\log p_\theta(z)\right]}_{R_z} + \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}\left[-\log p_\theta(x|z)\right]}_{D}\right], \tag{1}$$

which combines a distortion metric D and a rate term R_z that approximates the bitrate necessary to transmit the latent code z. Up to a constant entropy term $H[q_\phi]$, this loss equals the β-weighted VAE loss (Habibian et al., 2019; Higgins et al., 2017).

Much research has focused on the design of efficient network architectures for neural video codecs, in particular with respect to the temporal structure. While some earlier works were based on 3D convolutions (Habibian et al., 2019; Pessoa et al., 2020), recent models use optical flow and residual modeling to exploit similarities between frames. The latter class of approaches can be divided into predictive or P-frame temporal modeling, where the model for each frame x_t is conditional on the previous frame(s) (Agustsson et al., 2020; Chen et al., 2020; Liu et al., 2020; Lombardo et al., 2019; Lu et al., 2019; Rippel et al., 2019; Golinski et al., 2020; Rippel et al., 2021), and bidirectional or B-frame modeling, where each frame is conditioned on past as well as future reference frames (Cheng et al., 2019; Choi & Bajić, 2019; Djelouah et al., 2019; Wu et al., 2018; Park & Kim, 2019; Pourreza & Cohen, 2021). Agustsson et al. (2020) generalize optical flow to scale-space flow by adding dynamic blurring to the warping operation, improving the modeling of uncertainty and leading to state-of-the-art results.

Instance-adaptive compression Rate-distortion autoencoders are trained by minimizing the RD loss in Eq. (1) over a training dataset $\mathcal{D}$. This approach relies on the assumptions that the model is not only able to fit the training data well, but also generalizes to unseen datapoints at inference time. In practice, however, finite training data, limited model capacity, optimization imperfections, and domain shift — differences between train and test distributions — can degrade the performance.

In the context of image and video compression, several solutions have been proposed for this problem. For example, Campos et al. (2019) directly optimize the latents z for each datapoint x at encoding time. This alleviates the generalization problem for the encoder, but it does not solve it for the decoder and prior.

Since both the prior and the decoder need to be known on the receiver side, any update to these models based on test data needs to be signalled in the bitstream. Depending on the implementation,

this may increase the bitrate substantially. Some works tackle this problem by only updating small parts of the decoder-side models (Klopp et al., 2020; Lam et al., 2019; 2020; Zou et al., 2020), for instance only the biases or a small artifact removal network, but the RD gains of these methods are limited.

Rozendaal et al. (2021) suggest a method to adapt the *full* model to a *single* datapoint. The key idea is that the parameter *updates* $\delta = \theta - \theta_{\mathcal{D}}$, where $\theta_{\mathcal{D}}$ are the global model parameters resulting from training on $\mathcal{D}$, can be be transmitted very efficiently. After discretization, the quantized updates $\bar{\delta}$ are entropy-coded under a parameter update prior $p[\bar{\delta}]$ that assigns high probability mass (and thus low transmission cost) to zero-updates $\bar{\delta} = 0$. This procedure enables a neural codec in which the full model — encoder, decoder, and prior — is finetuned on each data sample at test time by minimizing the loss

$$\mathcal{L}_{\text{InstA}}(\phi, \delta) = \mathcal{L}_{\text{RD}}(\phi, \theta_{\mathcal{D}} + \bar{\delta}) + \beta \underbrace{\left(-\log p(\delta)\right)}_{R_\theta}. \tag{2}$$

The R_θ term reflects the increased bitstream length from coding the parameter updates $\bar{\delta}$. In Rozendaal et al. (2021), this approach was successfully demonstrated on I-frame compression, essentially compressing a set of images with similar content. In this work we adapt it to the compression of full video sequences, as we will describe in the following section.

3 INSTANCE-ADAPTIVE VIDEO COMPRESSION

We now introduce our instance-adaptive compression codec for full video codecs, or InstA for short. The key idea is to optimize the parameters of a rate-distortion VAE *for each video sequence to be transmitted*, and to send the relevant network parameters in a rate-efficient way to the decoder. This approach can be beneficial whenever a standard rate-distortion autoencoder would not generalize well, for instance because of limited training data or domain shift. In addition, instance adaptation allows us to use smaller models while maintaining most of the compression performance. While encoding the network parameters to the bitstream increases the length of the code, we describe a compression scheme in which this cost is negligible, especially when amortized over many frames of a video sequence.

Our approach is agnostic about the base model and we demonstrate it on two different architectures: a scale-space flow model in the low-latency setting and the B-EPIC model in the B-frame setting. In the following, we first describe the base models, before describing how InstA compresses and decompresses video sequences.

Low-latency base model: Scale-space flow (SSF) We first focus on the widely used *low-latency setting*, which we define as having only access to current or previous frames when decoding a frame. As a base model for this setting, we use the scale-space-flow architecture introduced by Agustsson et al. (2020). Any video sequence is first split into *groups of pictures* (GoP). The first frame in each GoP is modeled as an image without any dependency on previous frames, i. e. as an I-frame. All other frames are P-frames, modeled as

$$x_i = \text{Scale-Space-Warp}(x_{i-1}, g_i) + r_i\,, \tag{3}$$

where x_{i-1} is the reconstructed previous frame, g_i is the estimated scale-space flow field, r_i is the estimated residual, and the scale-space warping operation performs optical-flow warping of the previous frame with a dynamic, position-dependent amount of Gaussian blur. The I-frame images, P-frame scale-space flow g, and P-frame residuals r_i are compressed with separate *hyperprior* models (Ballé et al., 2018), a hierarchical variational autoencoder architecture.

We largely follow the architecture choices of Agustsson et al. (2020) and refer to this model as SSF18 after its number of decoder-side parameters (in millions). In Appendix B we describe the architecture in detail and highlight the differences to Agustsson et al. (2020).

Smaller low-latency base models As our instance-adaptive approach is based on finetuning the compression model on a low-entropy "dataset" (a single video sequence), we hypothesize that it does not require the full expressivity of the computationally complex SSF18 model. We propose three alternative scale-space flow architectures with a reduced computational complexity. In Tbl. 1

Table 1: Parameter counts and decoding complexity for our various global scale-space-flow models.

Model	SSF18	SSF8	SSF5	SSF3
Total parameters [10^6]	28.9	13.8	8.1	4.9
Decoder-side params. [10^6]	18.0	8.4	5.0	3.1
Decoder-side kMACs / pixel	313.4	180.0	88.6	50.8
Decoder kMACs reduction	0 %	43 %	72 %	84 %

we list the parameter counts and the number of multiply-accumulate (MAC) operations required for decoding a sequence, which are reduced by 43–84% compared to the SSF18 baseline. The architectures of the four SSF models are described in detail in Appendix B.

B-frame base model: B-EPIC We also consider a less constrained setting, in which frames can be compressed as B-frames as well, i. e. using both a previous frame and a future frame as reference points. This flexibility allows for even more efficient compression and is used for instance for on-demand video streaming. As a base model we choose the B-EPIC architecture (Pourreza & Cohen, 2021). Again, the video is split into GoPs. The very first frame of a video is modeled as an I-frame, while the last frame in each GoP is a P-frame, using the last frame of the preceeding GoP as reference; both I-frame and P-frames are compressed as in the SSF model.

Any other frame x_i is modeled as a B-frame: it is assigned a past reference frame x_j ($j < i$) and a future reference frame x_k ($k > i$); an off-the-shelf Super-SloMo frame interpolator (Jiang et al., 2018) is used to interpolate between these two reference frames. The interpolated frame is then used as a basis for scale-space flow warping:

$$x_i = \text{Scale-Space-Warp}(\text{Super-SloMo}(x_j, x_k), g_i) + r_i \,. \tag{4}$$

The interpolated frame provides a more useful starting point for optical-flow warping than the previous frame, for instance because the combination of past and future knowledge can often avoid occlusion effects.

Again, I-frames, optical flow, and residuals are modeled with three separate hyperprior models and the model is trained on the RD loss in Eq. (1). We only consider a single configuration and use hyperparameters and checkpoints from Pourreza & Cohen (2021). This model has 38.5 million parameters, 23.2 million of which are relevant on the decoder side.

Encoding a video sequence Our procedure conceptually follows Rozendaal et al. (2021). For completeness we shall describe the full method here. A video sequence x is compressed by:

1. Finetuning the model parameters (θ, ϕ) of the base model on the sequence x using Eq. (2),
2. computing the latent codes $z \sim q_\phi(z|x)$,
3. parameterizing the finetuned decoder and prior parameters as updates $\delta = \theta - \theta_\mathcal{D}$,
4. quantizing latent codes z and network updates δ, and
5. compressing the quantized latents $\bar{z}$ and updates $\bar{\delta}$ with entropy coding to the bitstream.

To finetune a pretrained base compression model, we start with the global model with parameters $(\theta_\mathcal{D}, \phi_\mathcal{D})$ and minimize the rate-distortion loss given in Eq. (2) — but only over the single video sequence x. This modified rate-distortion loss explicitly includes the bitrate required to send model updates δ under an update prior $p(\delta)$. We compute the regularizing R_θ loss with the unquantized updates δ, but use the quantized parameter updates $\theta_\mathcal{D} + \bar{\delta}$ to calculate the D and R_z loss terms, using a straight-trough estimator (Bengio et al., 2013) in the backward pass.

Sending the updated network parameters of course adds to the length of the bitstream. This is even the case when finetuning does not lead to changed parameters ($\bar{\delta} = 0$). Given the large size of the neural model we consider, it is therefore essential to choose an update prior that assigns a large probability mass to zero-updates $\bar{\delta} = 0$. This allows the network to transmit trivial updates at a negligible rate cost, while giving it the freedom to invest bit cost in non-trivial parameter updates that improve the performance substantially. We use a *spike-and-slab* prior (Johnstone & Titterington, 2009; Rozendaal et al., 2021), a mixture model of a narrow and a wide Gaussian distribution given by

$$p(\delta) = \frac{\mathcal{N}(\delta|0, \sigma^2 \mathbb{1}) + \alpha\, \mathcal{N}\big(\delta\big|0, s^2 \mathbb{1}\big)}{1 + \alpha} \,, \tag{5}$$

where the "slab" component with variance σ^2 keeps the bitrate cost for sizable updates down, and the "spike" component with the narrow standard deviation $s \ll \sigma$ ensures cheap zero-updates. The mixing weight α is a tunable hyperparameter.

At the beginning of the finetuning procedure, our neural model is equal to the global model. Because of the spike-slab update prior, the rate cost is only marginally increased and the compression performance essentially equal to the global model. During finetuning, the rate-distortion performance gradually improves, giving us an anytime algorithm that we can stop prematurely to get the best compression performance within a given encoder compute budget.

After the model has converged or a compute budget has been exhausted, we use the finetuned encoder q_ϕ to find the latent code z corresponding to the video sequence x. Both this latent code z and the updates to prior and decoder δ, which will be necessary to decode the video sequence, are quantized. To discretize the updates δ, we use a fixed grid of n equal-sized bins of width t centered around $\delta = 0$. We choose n such that $p(-n/t < \delta < n/t) \approx 1 - \epsilon$ and clip values at the tails. The quantization of z is analogous, except that we use a bin width of $t = 1$ and do not clip the values at the tails (in line with Ballé et al. (2018)).

Finally, we write the quantized updates $\bar{\delta}$ and quantized codes $\bar{z}$ to the bitstream. We use entropy coding under the quantized update prior $p(\bar{\delta})$ and finetuned prior $p_{\theta_\mathcal{D}+\bar{\delta}}(\bar{z})$, respectively. For latent codes in the tail region we use Exp-Golomb coding (Wiegand et al., 2003).

Decoding a video sequence The receiver first decodes the prior and decoder updates $\bar{\delta}$ from the bitstream. Once $\bar{\theta} = \theta_\mathcal{D} + \bar{\delta}$ is known on the decoder side, the latents z are decoded with the prior $p_{\bar{\theta}}(z)$ and the video sequence is reconstructed with the decoder $p_{\bar{\theta}}(x|z)$ following standard VAE protocol. The only overhead on the decoder side is therefore the initial decoding of $\bar{\delta}$. While this results in a small delay before the first frame can be decoded, in practice this delay is very short (below 0.2 seconds in our experiments).

4 Experimental setup

Datasets In our experiments we use sequences in Full-HD resolution (1920×1080 pixels) from four different datasets. The global models are trained on Vimeo90k (Xue et al., 2019). We evaluate on the HEVC class-B test sequences (HEVC, 2013), on the UVG-1k (Mercat et al., 2020) dataset, and on Xiph-5N (Rozendaal et al., 2021), which entails five Xiph.org test sequences. In addition, we evaluate the performance on out-of-distribution data on two sequences from the animated short film Big Buck Bunny, also part of the Xiph.org collection (Xiph.org). We choose an "easy" (BBB-E) and a "hard" (BBB-H) clip of 10s each. We describe how these are selected and all datasets in more detail in Appendix A.

Global models The scale-space flow models described in Sec. 3 are trained with the MSE training setup described in Agustsson et al. (2020), except that we use the publicly available Vimeo-90k dataset. For the B-EPIC model, we use the model trained by Pourreza & Cohen (2021) on Vimeo-90k. We describe the training schemes in detail in Appendix C.

Instance-adaptive finetuning On each instance, we finetune the models with the InstA objective in Eq. 3, using the same weight β as used to train the corresponding global model. In the low-latency scenario we use a GoP size of 3 and finetune on full-resolution frames (1920×1080 pixels) with a batch size of 1 and a learning rate of 10^{-5}. After finetuning, we transmit sequences with a GoP size of 12. The prior and decoder updates $\delta = \theta - \theta_\mathcal{D}$ are quantized and coded under the spike-and-slab network prior described in Sec. 3.

While the SSF model follows a stationary distribution over frames, the B-EPIC model has a more complex frame structure: the distance between a frame and its reference points can vary between 1 to 6 frames. For this reason, we found it beneficial to use a GoP of 12 for InstA finetuning of the B-EPIC model. Due to memory limitations, we finetune on horizontal crops of size 256×1080 pixels instead of full-resolution frames. For the low-bitrate working points use a learning rate of 10^{-5}, in line with Pourreza & Cohen (2021). For the high-bitrate region ($\beta \leq 0.0008$), we found that we can get faster convergence by using a learning rate of $5 \cdot 10^{-5}$.

Baselines As our instance-adaptive video compression method can be applied to any compression architecture, we focus on comparing to the respective base models: we compare InstA-SSF to our reimplementation of the SSF model (Agustsson et al., 2020) and InstA-B-EPIC to the B-EPIC model (Pourreza & Cohen, 2021). To better understand where the benefits of instance-adaptive finetuning come from, we consider an additional SSF18 baseline in which only the encoder is finetuned on each instance. Unlike our InstA method, encoder-only finetuning does not require sending any model updates in the bitstream. As additional neural baselines we show the results from Wu et al. (2018); Rippel et al. (2021); Lu et al. (2019).

We also run the popular classical codecs H.264 (AVC) (Wiegand et al., 2003) and H.265 (HEVC) (Sullivan et al., 2012) in the ffmpeg (x264 / x265) implementation (FFmpeg; VideoLAN, a;b) as well as the (slower but more effective) HM reference implementation (HEVC) of H.265. To ensure a fair comparison, in the low-latency experiments we restrict all codecs to I-frames and P-frames and fix the GoP size to 12. In the B-frame experiments, we use default settings, allowing the codecs to freely determine frame types and GoP size. For ffmpeg, we compare several different encoder presets to study the trade-off between encoding time and RD performance. In Appendix D we provide the exact commands used to generate these baseline results.

Metrics We evaluate the fidelity of the reconstructions through the peak signal-to-noise ratio (PSNR) in RGB space. To summarize the rate-distortion performance in a single number, we compute the Bjøntegaard Delta bitrate (BD-rate) (Bjøntegaard, 2001), a measure of the rate difference between two codecs averaged over a range of distortion values. We evaluate the BD-rate savings of all methods relative to the HM reference implementation of HEVC, using PSNR as distortion metric, the intersection of the support of all RD curves as the distortion range, and cubic spline interpolation.

5 RESULTS

Compression performance Figure 1 shows the rate-distortion curves of our instance-adaptive video codec (InstA) as well as neural and traditional baselines. Both for SSF in the low-latency and B-EPIC in the B-frame setting, the instance-adaptive models clearly outperform the corresponding base models. Finetuning only the encoder — which does not require sending model updates in the bitstream — leads to a much more modest improvement over the base models.

In the low-latency setting, InstA-SSF outperforms all other neural models as well as the ffmpeg implementations of H.264 and H.265 at all studied bitrates, except for ELF-VC, which holds an edge in the low-bitrate region. Within the restrictions of the low-latency setting and at medium to high bitrates, it is even competitive with the HM reference codec. In the B-frame configuration, InstA-B-EPIC outperforms all baselines except HM from around 0.1 bits per pixel.

Robustness to model-size reduction While the full-size InstA-SSF18 model leads to the best RD performance in the low-latency setting, the smaller InstA-SSF5 models hold their own. Despite the 72% reduction in decoder complexity, Insta-SSF5 outperforms all neural baselines and the ffmpeg codecs, and achieves a performance close to that of InstA-SSF18. In comparison, the gap between the SSF5 and SSF18 baselines is larger.

To further investigate how the performance of our method depends on decoder complexity, we train two additional models, SSF3 and SSF8, as summarized in Tbl. 1. The results are in line with the trends already shown. While the performance of the global models degrades quickly when reducing the model size, instance-adaptive finetuning can compensate for the reduced capacity and offers more "bang for the buck" in terms of (decoding-side) model complexity. For detailed results we refer the readers to Appendix G.

Robustness to dataset variation Globally trained neural codecs rely on the similarity of the test instances to the training data. In comparison, we expect our finetuned models to be more robust under domain shift and to perform well across different test sets. To put this to the test, we evaluate the InstA-SSF models and low-latency baselines on different sequences, including two animated scenes from the Big Buck Bunny video. In Tbl. 2 we report the compression performance as BD-rate savings relative to the HM reference codec, the corresponding RD curves can be found in

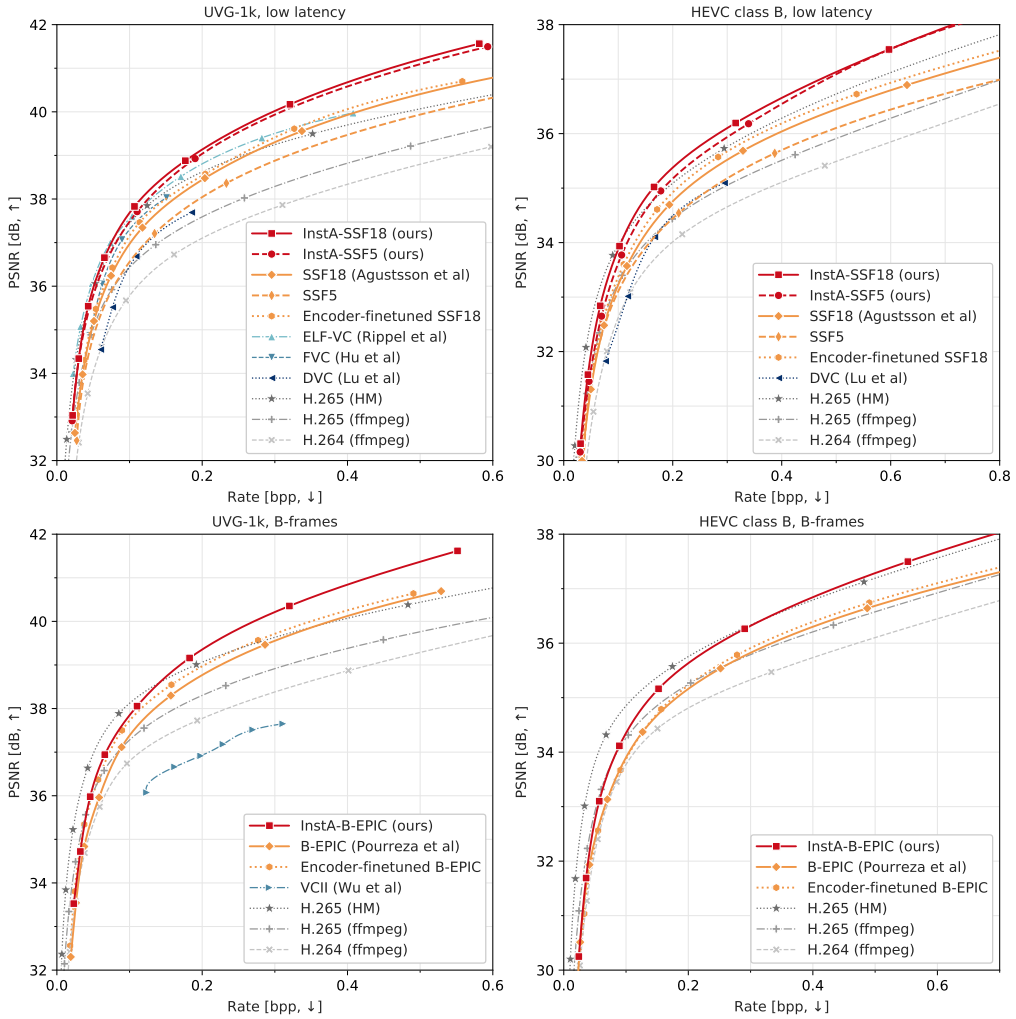

Figure 1: Rate-distortion performance of our InstA video codecs. InstA (red) leads to BD-rate savings of between 17 % and 26 % over the corresponding base models (orange) and between 6 % and 44 % over H.265 (ffmpeg).

Appendix E. We find that InstA performs strongly on all datasets. On the animated sequences, the global SSF18 baseline performs poorly relative to the ffmpeg H.265 codec, showing its susceptibility to domain shift. Our instance-adaptive models are much more robust to this shift and close the gap with classical codecs. We also illustrate this performance improvement on example frames from the Big Buck Bunny video in Appendix F and provide reconstructed video data in the supplementary material.

To further quantify our robustness to test data characteristics, we investigate the effect of the video framerate on the model performance. In Appendix G we study the performance of global and instance-adaptive SSF models on temporally subsampled videos and again find that the instance-adaptive model is more robust to variations than the global SSF model. While these ablations do not cover all realistic variations, they serve as examples of dataset characteristics that a global neural codec is likely sensitive to.

Rate composition It is remarkable that InstA consistently improves the rate-distortion performance of neural models despite having to send updates for millions of neural network parameters. We investigate this low rate overhead in Appendix G. We find that the spike-and-slab prior is crucial: compared to Gaussian coding, it allows us to save around 7.5 bits per parameter.

Table 2: Rate-distortion performance of InstA and baselines on several datasets. We summarize the rate-distortion performance as BD-rate relative to the HM reference implementation of H.265 in the respective setting (lower is better).

Setting	Method	BD-rate relative to H.265 (HM) [↓]				
		UVG-1k	HEVC-B	Xiph-5N	BBB-E	BBB-H
Low latency	InstA-SSF18 (ours)	$\mathbf{-11.0}\,\%$	$\mathbf{4.7}\,\%$	$\mathbf{-10.4}\,\%$	$\mathbf{13.4}\,\%$	$28.3\,\%$
	InstA-SSF5 (ours)	$-6.4\,\%$	$13.7\,\%$	$-7.5\,\%$	$31.9\,\%$	$43.2\,\%$
	SSF18 Agustsson et al. (2020)	$20.4\,\%$	$37.3\,\%$	$17.4\,\%$	$32.7\,\%$	$54.9\,\%$
	SSF5	$41.4\,\%$	$51.2\,\%$	$41.4\,\%$	$84.5\,\%$	$99.3\,\%$
	Encoder-finetuned SSF‘8	$14.5\,\%$	$28.2\,\%$			
	H.265 (ffmpeg)	$58.5\,\%$	$47.1\,\%$	$29.5\,\%$	$25.9\,\%$	$\mathbf{18.7}\,\%$
	H.264 (ffmpeg)	$113.1\,\%$	$92.7\,\%$	$69.3\,\%$	$108.4\,\%$	$46.9\,\%$
B-frames	InstA-B-EPIC (ours)	$\mathbf{26.2}\,\%$	$\mathbf{44.7}\,\%$			
	B-EPIC (Pourreza & Cohen, 2021)	$57.4\,\%$	$74.9\,\%$			
	Encoder-finetuned B-EPIC	$36.7\,\%$	$75.9\,\%$			
	H.265 (ffmpeg)	$61.0\,\%$	$53.1\,\%$			
	H.264 (ffmpeg)	$118.4\,\%$	$107.0\,\%$			

InstA learns to distribute bits on model updates and latents. The optimal trade-off generally depends on the bitrate, the size of the model and latent code, and on the test data: we expect that for strongly out-of-distribution data, more bits will be spent on model updates. Indeed we find that in almost all cases, less than 1% of the total rate is spent on model updates, while for the animated sequences from the Big Buck Bunny video between 1% and 3% of the total bitrate are invested in model updates. This confirms the intuition that the InstA codec adapts to domain shift by more strongly changing the prior and decoder model.

Computational complexity The improved compression performance of our instance-adaptive video codec comes at the price of increased encoding time. In Fig. 2 we illustrate this trade-off for the low-latency setting. We summarize the rate-distortion performance as BD-rate savings relative to HM and compare to the encoding time[1]. For our instance-adaptive methods, this time includes the finetuning in addition to the forward pass of the encoder, the quantization, and the entropy coding (for both the latents and the parameters). By choosing how long to run the finetuning, we can trade off RD performance against the encoding time. Out of the studied codecs, ours can best exploit this trade-off: encoding time and compression performance are fixed in other neural methods, while the existing handles in traditional codecs have a limited effect on performance.

We find that instance-adaptive finetuning leads to state-of-the-art RD performance even when finetuning only for a few steps. The performance gradually improves over the course of training and eventually flattens out. Care has to be taken in interpreting these results as the different codecs were designed for different hardware and our algorithms have not been optimized for performance.

Instance-adaptive video compression also adds some overhead on the decoder side, as the model parameters need to be decoded before any frame can be reconstructed. For the full-size InstA-SSF18 model and without optimizing the code for performance, this initial delay is below 0.2 seconds. The neural network forward pass alone is timed with a speed of around 0.01 seconds per frame on a TeslaV100 GPU without any optimization. Both the overhead and the network forward pass are thus negligible compared to the time it takes to entropy-decode the latents z: in our (not optimized) implementation, this takes around 1 second per frame, both for the global SSF models and InstA-SSF. Compared to the neural baselines, InstA-SSF thus yields a better performance at comparable complexity. Compared to traditional codecs, there is still a gap to bridge, and none of the neural codecs can decode at the framerate of the videos yet.

At the same time, instance-adaptive finetuning allows us to use smaller architectures while maintaining a strong performance, as shown above. This potential to reduce the decoder-side compute may outweigh the small overhead from the initial update decoding.

[1]We report the walltime on machines with 40-core Intel Xeon Gold 6230 CPUs with 256 GB RAM and NVIDIA Tesla V100-SXM2 GPUs with 32 GB VRAM. We only use a single GPU.

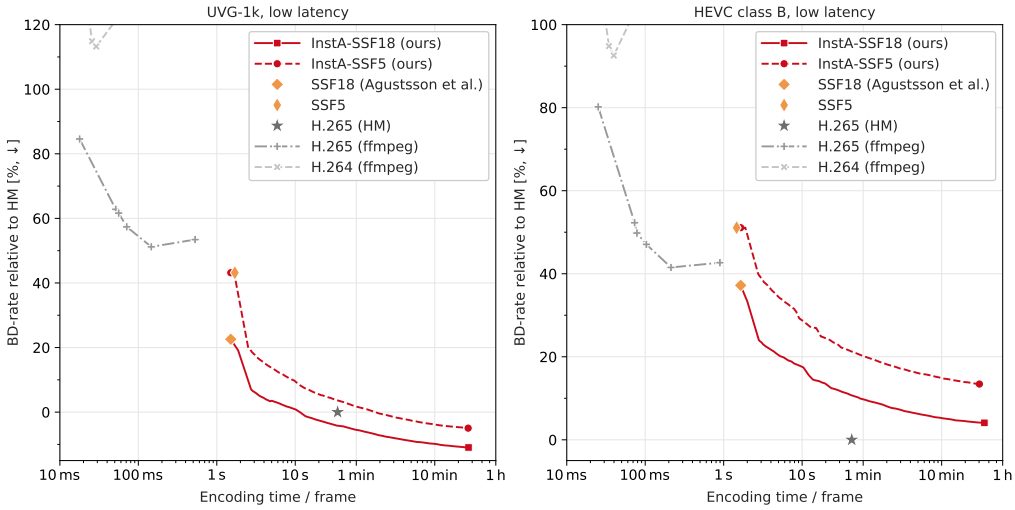

Figure 2: Trade-off between RD performance and encoder-side compute in the low-latency setting. We summarize the RD performance as BD-rate difference relative to the HM implementation of H.265 (lower is better). InstA-SSF (red) is compared to SSF (orange) and traditional (grey) baselines. Some fluctuations in the measured walltime are due to varying load on our computational cluster. Since we can stop the instance finetuning at any time, InstA-SSF allows us to continuously trade off encoder compute for compression performance.

Model update strategy for longer sequences For longer videos with multiple scenes, one can apply instance-adaptive finetuning per scene, per video, or in an hierarchical approach. In Appendix G we discuss these aspects and compare them for the Big Buck Bunny video.

6 Conclusions

We introduced InstA, a method for video compression that finetunes a pretrained neural compression model on each video sequence to be transmitted. The finetuned network parameters are compressed and transmitted along with the latent code. The method is general and can be applied to many different settings and architectures, we demonstrated it on scale-space flow models in the low-latency setting as well as a state-of-the-art B-frame architecture.

Instance-adaptive video compression relaxes the requirement of a single global model that needs to be able to generalize to any unseen test sequence. This can be beneficial when the training data is limited, when there is domain shift between training and test distributions, or when the model class is not expressive enough. While it comes at the cost of some overhead to the bitstream, we demonstrated that with a suitable compression scheme this overhead can be kept small, especially since it is amortized over the many pixels of a video sequence.

Our InstA models clearly outperform the base models in all considered settings, in several cases achieving a state-of-the-art rate-distortion performance. They outperform their corresponding global base models with BD-rate savings of between 17 % and 26 % on the UVG-1k, HEVC class-B, and Xiph-5N datasets, and the popular ffmpeg implementation of H.265 with BD-rate savings between 6 % and 44 %. By evaluating models that were trained on natural scene data on animated sequences, we demonstrate that instance-adaptive finetuning improves the robustness of neural video codecs under domain shift. Finally, instance-adaptive compression reduces the required model expressivity: even after reducing the model size by 72 %, InstA-SSF maintains a state-of-the-art performance.

The ability to trade off encoder-side compute for compression performance is a natural match for one-to-many transmission scenarios, in which a sequence is compressed once but transmitted and decoded often. We demonstrated that even just finetuning for seconds per frame can lead to a substantial improvement in compression performance, so instance-adaptive compression may ultimately even be beneficial in real-time applications.

REFERENCES

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*, pp. 8503–8512, 2020.

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.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Gisle Bjøntegaard. Calculation of average psnr differences between rd-curves (vceg-m33). In *VCEG Meeting (ITU-T SG16 Q. 6)*, pp. 2–4, 2001.

Benjamin Bross, Jianle Chen, Shan Liu, and Ye-Kui Wang. Versatile video coding (draft 5). *Joint Video Experts Team (JVET) of ITU-T SG*, 16:3–12, 2018.

Joaquim Campos, Simon Meierhans, Abdelaziz Djelouah, and Christopher Schroers. Content adaptive optimization for neural image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

Z. Chen, T. He, X. Jin, and F. Wu. Learning for video compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2):566–576, 2020.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learning image and video compression through spatial-temporal energy compaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Hyomin Choi and Ivan V Bajić. Deep frame prediction for video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7):1843–1855, 2019.

Abdelaziz Djelouah, Joaquim Campos, Simone Schaub-Meyer, and Christopher Schroers. Neural Inter-Frame compression for video coding. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6421–6429, 2019.

FFmpeg. FFmpeg. `http://ffmpeg.org/`.

Adam Golinski, Reza Pourreza, Yang Yang, Guillaume Sautiere, and Taco S Cohen. Feedback recurrent autoencoder for video compression. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Amirhossein Habibian, Ties van Rozendaal, Jakub M Tomczak, and Taco S Cohen. Video compression with rate-distortion autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7033–7042. openaccess.thecvf.com, 2019.

HEVC. HEVC test model (HM). `https://vcgit.hhi.fraunhofer.de/jvet/HM`.

HEVC. Common test conditions and software reference configurations. `http://phenix.it-sudparis.eu/jct/doc_end_user/current_document.php?id=7281`, 2013.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. *5th International Conference on Learning Representations (ICLR)*, 2017.

Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9000–9008, 2018.

I. Johnstone and D. Titterington. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367: 4237 – 4253, 2009.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2014.

Jan P Klopp, Liang-Gee Chen, and Shao-Yi Chien. Utilising low complexity cnns to lift non-local redundancies in video coding. *IEEE Transactions on Image Processing*, 2020.

Yat-Hong Lam, Alireza Zare, Caglar Aytekin, Francesco Cricri, Jani Lainema, Emre Aksu, and Miska Hannuksela. Compressing weight-updates for image artifacts removal neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

Yat-Hong Lam, Alireza Zare, Francesco Cricri, Jani Lainema, and Miska M Hannuksela. Efficient adaptation of neural network filter for video compression. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 358–366, 2020.

Haojie Liu, Han Shen, Lichao Huang, Ming Lu, Tong Chen, and Zhan Ma. Learned video compression via joint spatial-temporal correlation exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11580–11587, 2020.

Salvator Lombardo, Jun Han, Christopher Schroers, and Stephan Mandt. Deep generative video compression. In *Advances in Neural Information Processing Systems 32*, pp. 9287–9298. Curran Associates, Inc., 2019. URL `http://papers.nips.cc/paper/9127-deep-generative-video-compression.pdf`.

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*, pp. 11006–11015, June 2019.

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*, pp. 297–302, 2020. URL `http://ultravideo.fi/#testsequences`.

Woonsung Park and Munchurl Kim. Deep predictive video compression with bi-directional prediction. *arXiv preprint arXiv:1904.02909*, 2019.

Jorge Pessoa, Helena Aidos, Pedro Tomás, and Mário AT Figueiredo. End-to-end learning of video compression using spatio-temporal autoencoders. In *2020 IEEE Workshop on Signal Processing Systems (SiPS)*, pp. 1–6. IEEE, 2020.

Reza Pourreza and Taco S Cohen. Extending neural p-frame codecs for b-frame coding. *arXiv preprint arXiv:2104.00531*, 2021.

Oren Rippel, Sanjay Nair, Carissa Lew, Steve Branson, Alexander G Anderson, and Lubomir Bourdev. Learned video compression. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3454–3463, 2019.

Oren Rippel, Alexander G Anderson, Kedar Tatwawadi, Sanjay Nair, Craig Lytle, and Lubomir Bourdev. ELF-VC: Efficient Learned Flexible-Rate Video Coding. *arXiv preprint arXiv:2104.14335*, 2021.

Ties van Rozendaal, Iris AM Huijben, and Taco S Cohen. Overfitting for fun and profit: Instance-adaptive data compression. In *International Conference on Learning Representations (ICLR)*, 2021.

Sandvine. 2019 Global Internet Phenomena Report. `https://www.ncta.com/whats-new/report-where-does-the-majority-of-internet-traffic-come`, 2019.

Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012.

Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*, March 2017.

VideoLAN. x264 library. https://www.videolan.org/developers/x264.html, a.

VideoLAN. x265 library. https://www.videolan.org/developers/x265.html, b.

Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the H.264/AVC video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7): 560–576, 2003.

Chao-Yuan Wu, Nayan Singhal, and Philipp Krahenbuhl. Video compression through image interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 416–431, 2018.

Xiph.org. Video test media. https://media.xiph.org/video/derf/.

Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127(8):1106–1125, 2019. URL http://toflow.csail.mit.edu/.

Nannan Zou, Honglei Zhang, Francesco Cricri, Hamed R Tavakoli, Jani Lainema, Miska Hannuksela, Emre Aksu, and Esa Rahtu. L^2C–learning to learn to compress. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6. IEEE, 2020.

A DATASETS

UVG-1k Following Refs. (Wu et al., 2018; Lu et al., 2019; Golinski et al., 2020; Agustsson et al., 2020), we report results on the UVG-1k dataset, which consists of seven test video sequences totalling 3900 frames captured at 120 fps and natively available in 1080p resolution. Note that these sequences are only a subset of the full UVG dataset (Mercat et al., 2020).

Xiph-5N We use the Xiph-5N dataset introduced in Rozendaal et al. (2021). It consists of the following five sequences from the Xiph.org collection: "in_to_tree", "aspen", "controlled_burn", "sunflower", and "pedestrian_area".

Big Buck Bunny (BBB-E and BBB-H) From the Big Buck Bunny video, which is also part of the Xiph.org media collection, we extract two subsequences of 10 s length. They are chosen based on the frame-wise RD loss achieved by a global SSF18 model (with $\beta = 0.0016$). We select the 10 s window with the lowest RD loss and the 10 s window with the highest RD loss, excluding the intro and outro of the video. These signify the "easiest" and "hardest" parts of the video to compress and we refer to them as "BBB-E" and "BBB-H", respectively. In Fig. 3 we show the frame-wise RD loss and the selected sequences.

B MODEL ARCHITECTURES

Full-size SSF model We first apply instance-adaptive video compression to the scale-space flow (SSF) architecture proposed by Agustsson et al. (Agustsson et al., 2020), using our own re-implementation of the model. We mostly follow the setup described in Agustsson et al. (2020), but there are some minor differences. In the hyperencoder and hyperdecoder networks we use a slightly different architecture. We list the architectures in Tbl. 3. In addition, our implementation of the blur stack and of the scale-space-flow warping operation differ from the one used by Ref. 3, but we have verified in experiments that this has a negligible effect.

Smaller SSF models In addition to the full-size SSF18 model, we consider three smaller scale-space flow architectures. They follow the same structure as SSF18 in Tbl. 3 but differ in the number of latent variables and convolutional channels, as shown in Tbl. 4. We refer to them as SSF8, SSF5, and SSF3 after the number of decoder-side network parameters (in millions).

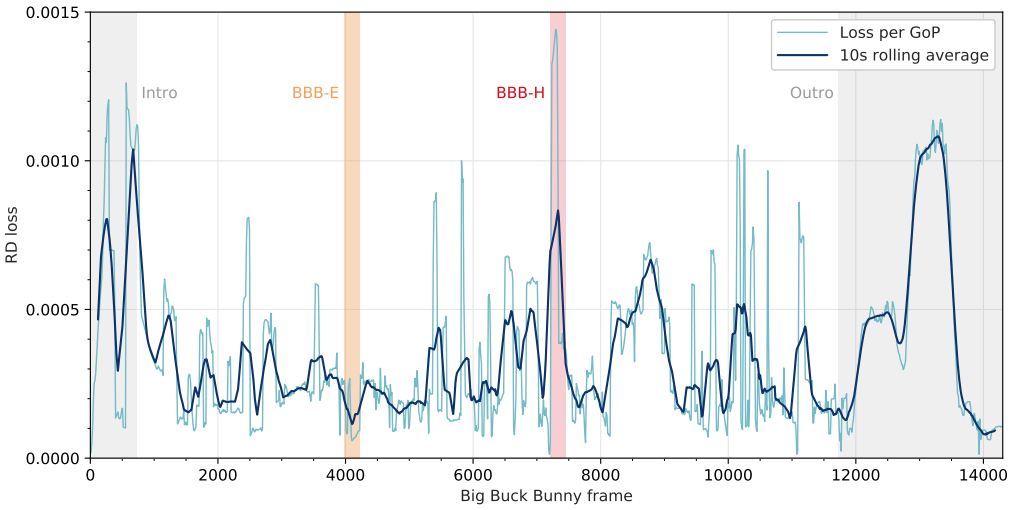

Figure 3: Temporal distribution of the RD loss of the global SSF18 model for the Big Buck Bunny video. We use the 10 s windows with the lowest and highest loss as benchmarks, excluding the intro and outro of Big Bunny. These subsequences, which we call "BBB-E" and "BBB-H" for "easy to compress" and "hard to compress", are marked in orange and red.

Table 3: Architecture differences between the original SSF model published by Agustsson et al. (2020) and our SSF18 re-implementation. Convolutional layers are indicated as "conv [kernel size] [layer type][stride] [output channels]". $\downarrow$ is used for convolutional layers, while $\uparrow$ is used for transposed convolutions. The hyperdecoder is modeled as a Gaussian $p(z_2|z_1) = \mathcal{N}(z_2|\mu(z_1), \sigma(z_1))$, where the mean and standard deviation are modeled by two separate neural networks $\mu(z_1)$ and $\sigma(z_1)$ with largely identical architectures. These network differ in their activation functions, activations that are only used in one of the two networks are indicated in the table by their respective symbol. Note that we always clamp $\sigma(z_1)$ to the range $[0.11, \infty)$.

Agustsson et al. (Agustsson et al., 2020)	Our reimplementation
Hyperencoder $q(z_1\|z_2)$	
conv 5x5 $\downarrow$2 192	conv 3x3 $\downarrow$1 192
ReLU	ReLU
conv 5x5 $\downarrow$2 192	conv 5x5 $\downarrow$2 192
ReLU	ReLU
conv 5x5 $\downarrow$2 192	conv 5x5 $\downarrow$2 192
Hyperdecoder $p(z_2\|z_1)$	
conv 5x5 $\uparrow$2 192	conv 5x5 $\uparrow$2 192
μ: ReLU, σ: QReLU	ReLu
conv 5x5 $\uparrow$2 192	conv 5x5 $\uparrow$2 192
μ: ReLU, σ: QReLU	ReLu
conv 5x5 $\uparrow$2 192	conv 3x3 $\uparrow$1 192
σ: QReLU	σ: ReLU

B-EPIC model As a second base model we use the B-frame codec developed by Pourreza & Cohen (2021) (B-EPIC). We follow the hierarchical IBP configuration presented in that paper and use the model trained by the authors. During training, a GoP consists of 4 frames arranged in an IBBP pattern. At test time, we use a GoP size of 12 frames and the following reference structure (where frame 0 refers to the last frame of the previous GoP):

- Frame 12: P-frame referring to frame 0
- Frame 6: B-frame referring to frames 0 and 12
- Frame 3: B-frame referring to frames 0 and 6
- Frame 9: B-frame referring to frames 6 and 12
- Frame 1: B-frame referring to frames 0 and 3
- Frame 2: B-frame referring to frames 1 and 3

and so on for the remaining frames. The very first frame in a video is encoded as an I-frame. Encoding and decoding in this order ensures that the reference frames will always be encoded or decoded before the frames for which they are needed.

Quantization settings Ablation studies showed insensitivity to network prior settings in the ranges of $10 \leq \alpha \leq 1000$, $0.0005 \leq t \leq 0.005$, and $0.01 \leq \sigma \leq 0.05$. For the experiments in this paper we use a bin width $t = 0.001$, $\sigma = 0.05$, $s = t/6$, a spike-slab ratio $\alpha = 100$, and $\epsilon = 2^{-8}$.

Entropy coding We implement an entropy coder for the latents and model updates in the SSF and InstA-SSF models. We verified that this achieves bitrates in agreement with the probability mass functions of $p(z)$ and $p(\delta)$. For the B-EPIC and InstA-B-EPIC models, we report bitrates based on probability mass functions without actually implementing an entropy coder.

C TRAINING SCHEME

Global models The scale-space flow models described in Sec. 3 are trained with the MSE training setup described in Agustsson et al. (2020), except that we use the publicly available Vimeo-90k

Table 4: Architecture details for our scale-space-flow models. In the top, we show the number of channels for the latent and hyperlatent variables, z, which are the basis for the compressed code; in the bottom, we show the number of channels in the codec and hypercodec networks in the encoder and decoder. We use the same model architecture and size for I-frame, scale-space-flow and residual networks.

Model	SSF18	SSF8	SSF5	SSF3
latent channels	192	192	192	128
hyperlatent channels	192	192	192	192
codec channels	128	96	64	48
hypercodec channels	192	96	64	48

dataset. The D loss is computed with (integer) quantized latents $\bar{z}$, while for the R_z loss term we use noisy samples from $U(z - \frac{1}{2}, z + \frac{1}{2})$. Following Agustsson et al. (2020), we first train for 1 million steps on 256×256 crops with a learning rate of 10^{-4}. We then conduct the MSE "finetune" stage of the training procedure from Agustsson et al. (2020) (not to be confused with instance-adaptive finetuning) for the SSF18 model, where we train on crops of size $h \times w = 256 \times 384$ with a learning rate of 10^{-5}. The models are trained with a GoP size of 3 frames, which means that we split the training video into chunks of 3 frames and randomly sample chunks during training. We finally evaluate the models with a GoP size of 12. This gives similar results to using infinite unrolling at test time, as we demonstrate in Appendix B.

For the B-EPIC model, we use the model trained by Pourreza & Cohen (2021) on Vimeo-90k. The setup is similar to that for the SSF models except that B-EPICs more complicated GoP structure requires training with a GoP size of 4 frames. At test time we use a GoP size of 12, the frame configurations are described in Appendix B.

D CLASSICAL CODECS

We generate H.265 and H.264 results using version v3.4.8 of ffmpeg (FFmpeg). When comparing to neural codecs in the low-latency setting, we also choose a low-latency setting that only allows I-frames and P-frames and fix the GoP size to 12. We use various encoder presets. Example commands are

```
ffmpeg -pix_fmt yuv420p -s 1920x1080 \
-r 120 -i raw_video.yuv \
-c:v libx264 -preset medium \
-crf 23 -tune zerolatency \
-x264-params \
"keyint=12:min-keyint=12:verbose=1" \
output.mkv
```

for H.264 and

```
ffmpeg -pix_fmt yuv420p -s 1920x1080 \
-r 120 -i raw_video.yuv \
-c:v libx265 -preset medium \
-crf 23 -tune zerolatency \
-x265-params \
"keyint=12:min-keyint=12:verbose=1" \
output.mkv
```

for H.265. We also compress the test sequences with the HM reference implementation of HEVC, using the "lowlatencyP" setting with a fixed GoP size of 12.

As baselines for the neural B-frame codecs, we also relax the GoP and frame-type restrictions for the classical codecs, using the default random-access settings with B-frames and variable GoP size.

Table 5: BD-rate of several codecs (rows) compared to different reference codecs (columns) in the low-latency setting. Lower numbers are better, the best results on each dataset are shown in bold.

Dataset	Method	BD-rate relative to…			
		H.265 (HM)	H.265 (ffmpeg)	H.264 (ffmpeg)	SSF18
UVG	InstA-SSF18 (ours)	**−11.0** %	**−43.8** %	**−58.2** %	**−26.1** %
	InstA-SSF5 (ours)	−6.4 %	−41.0 %	−56.1 %	−22.3 %
	SSF18	20.4 %	−24.0 %	−43.5 %	
	Encoder-finetuned SSF18	14.5 %	−27.8 %	−46.3 %	−4.9 %
HEVC-B	InstA-SSF18 (ours)	**4.7** %	**−28.8** %	**−45.7** %	**−23.8** %
	InstA-SSF5 (ours)	13.7 %	−22.7 %	−41.0 %	−17.2 %
	SSF18	37.3 %	−6.6 %	−28.7 %	
	Encoder-finetuned SSF18	28.2 %	−12.8 %	−33.4 %	−6.6 %
xiph5N	InstA-SSF18 (ours)	**−10.4** %	**−30.8** %	**−47.1** %	**−23.7** %
	InstA-SSF5 (ours)	−7.5 %	−28.6 %	−45.4 %	−21.3 %
	SSF18	17.4 %	−9.3 %	−30.7 %	
BBB-E	InstA-SSF18 (ours)	**13.4** %	**−9.9** %	**−45.6** %	**−14.5** %
	InstA-SSF5 (ours)	31.9 %	4.8 %	−36.7 %	−0.7 %
	SSF18	32.7 %	5.4 %	−36.3 %	
BBB-H	InstA-SSF18 (ours)	**28.3** %	**8.1** %	**−12.6** %	**−17.1** %
	InstA-SSF5 (ours)	43.2 %	20.6 %	−2.5 %	−7.5 %
	SSF18	54.9 %	30.5 %	5.4 %	

Table 6: BD-rate of several codecs (rows) compared to different reference codecs (columns) in the B-frame setting. Lower numbers are better, the best results on each dataset are shown in bold.

Dataset	Method	BD-rate relative to…			
		H.265 (HM)	H.265 (ffmpeg)	H.264 (ffmpeg)	B-EPIC
UVG-1k	InstA-B-EPIC (ours)	**26.2** %	**−21.7** %	**−42.2** %	**−19.9** %
	B-EPIC	57.4 %	−2.3 %	−27.9 %	
	Encoder-finetuned B-EPIC	36.7 %	−15.1 %	−37.4 %	−13.1 %
HEVC-B	InstA-B-EPIC (ours)	**44.7** %	**−5.5** %	**−30.1** %	**−17.3** %
	B-EPIC	74.9 %	14.2 %	−15.5 %	
	Encoder-finetuned B-EPIC	75.9 %	14.8 %	−15.0 %	0.5 %

E ADDITIONAL RATE-DISTORTION RESULTS

In Tbl. 5 we show the BD-rate savings of the InstA-SSF18, InstA-SSF5, and SSF18 models relative to two different H.265 implementations and relative to the SSF18 model. Figures 5 and 6 show rate-distortion curves on the Xiph-5N and Big Buck Bunny datasets.

F PERCEPTUAL QUALITY

We illustrate the performance on out-of-distribution data in Fig. 7, where we show original and compressed versions of two frames from the Big Buck Bunny video. Both traditional codecs and the SSF18 baseline lose details and introduce artifacts at the chosen low-bitrate setting. While the InstA-SSF18 compressions still suffer from some blurriness, artifacts are less pronounced than in the global model.

In the supplementary materials, we attach the corresponding reconstructed video sequences. To highlight the differences between the codecs, we again use a very low bitrate for all codecs (even though our InstA method is strongest at higher bitrates).

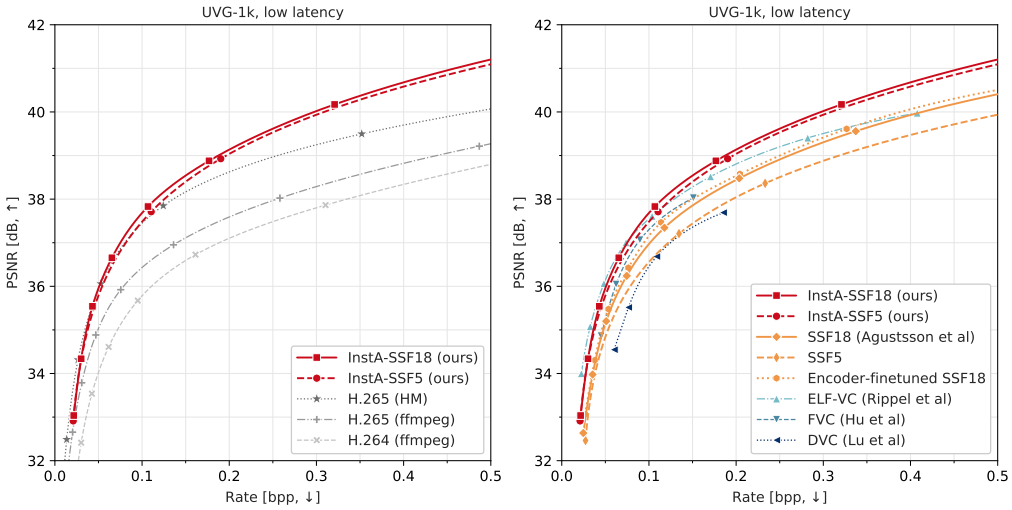

Figure 4: Rate-distortion performance of our instance-adaptive SSF models (InstA-SSF, red) on UVG in the low-latency setting. These are the same results as in the top left panel of Fig. (1), but separated into classical (left) and neural (right) baselines and zoomed in for better readability.

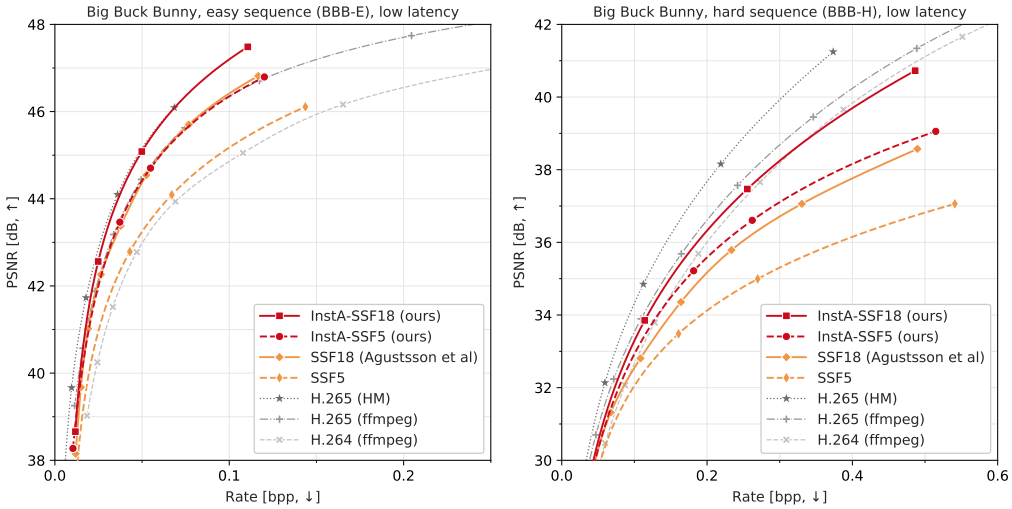

Figure 5: Rate-distortion performance of our instance-adaptive SSF models (InstA-SSF, red) compared to neural (orange) and traditional (grey) baselines on the “easy” (left) and “hard” (right) sequences from the Big Buck Bunny video.

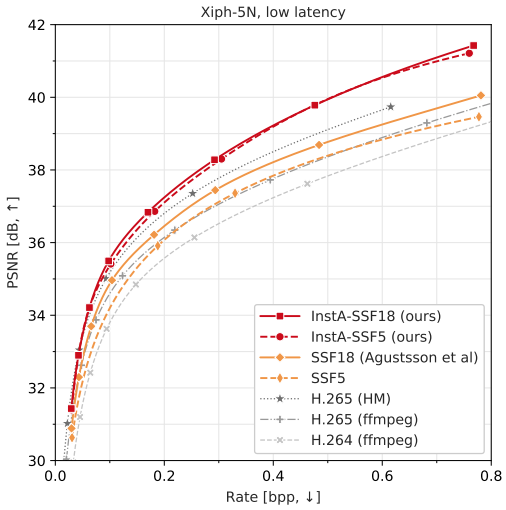

Figure 6: Rate-distortion performance of our instance-adaptive SSF models (InstA-SSF, red) compared to neural (orange) and traditional (grey) baselines on five Xiph.org test sequences.

Figure 7: Original and compressed details in frames 7395 (left) and 7276 (right) of the Big Buck Bunny video (Xiph.org). From top to bottom, we compare the original, compressed frames with the H.264 and H.265 codecs in the ffmpeg (x264 / x265) implementation, the SSF18 baseline, and our proposed InstA-SSF18 method. We use settings with comparable, low bitrates; the achieved bitrates (over the 10s sequence) are given in the left panels. All codecs lose textural details, but our InstA-SSF18 codec reduces the artifacts visible in the SSF18 reconstructions.

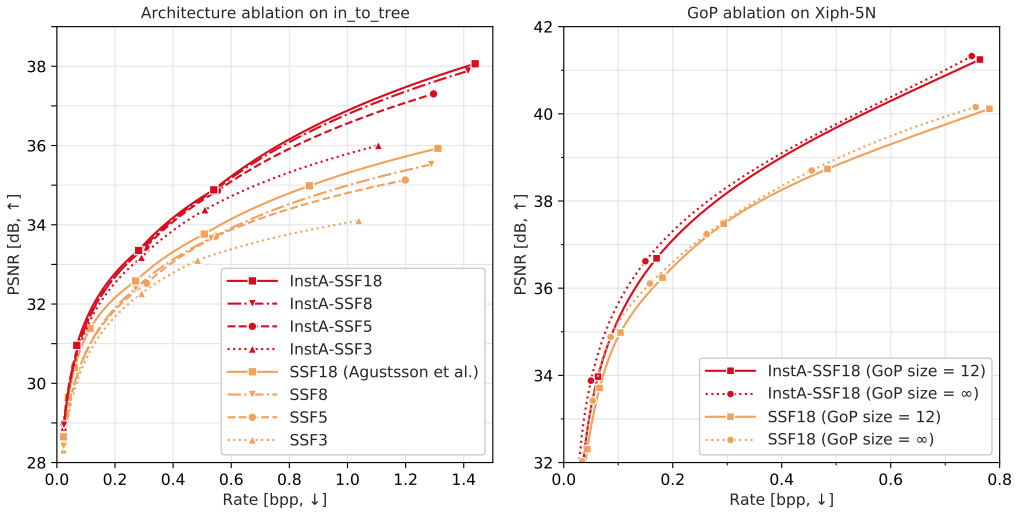

Figure 8: *Left:* Model size ablation study. We compare several global and instance-finetuned scale-space-flow models, showing the performance on the in_to_tree sequence (Xiph.org). *Right:* Effect of the GoP size at test time. We show rate-distortion curves of instance-adaptive and global SSF18 models on the HEVC Xiph-5N dataset, comparing a fixed GoP size of 12 (solid lines) to an infinite GoP size (dotted lines).

G Ablation Studies

Model size ablation In the left panel of Fig. 8 we compare the compression performance of globally trained SSF18, SSF8, SSF5, and SSF3 models, as well as instance-adaptive versions InstA-SSF18, InstA-SSF8, InstA-SSF5, and InstA-SSF3. Due to the training time requirement we limit this ablation to a single video, the "in_to_tree" sequence from the Xiph.org collection. We find that the larger models lead to a better compression performance, with more pronounced difference in the high-bitrate region. SSF3 and InstA-SSF3 models perform poorly at high bitrates. We attribute this partially to the small size of the latent space compared to all other models. In general, we observe that instance-adaptive training improves the performance of models of any complexity. Instance adaption reduces the gap between smaller models and the full-size model, confirming that less expressivity is necessary for good performance on a single instance compared to on the entire data distribution.

GoP size considerations Throughout this work, we use a GoP size of 12 at test time. To test the robustness of this choice, in the right panel of Fig. 8 we compare to an infinite GoP size (for the SSF model). In other words, we compress the first frame as I-frame and all remaining frames as P-frames. The overall difference is slim, though the gap is larger for some videos than others.

Sensitivity to framerate variations To further quantify our robustness to test data characteristics, we investigate the effect of the video framerate on the model performance. We study the performance of global and instance-adaptive SSF models on temporally subsampled videos, i. e. keeping one out of every four frames. To save computational resources, this ablation is only performed on the "Bosphorus" sequence from the UVG-1k dataset. The left panel of Fig. 9 shows our results. While the SSF18 performance varies greatly between the original and the subsampled video, the difference in performance is much smaller for InstA-SSF18. In addition, both models perform better in the 120 fps scenario due to the higher redundancy across frames.

Update sparsity analysis The key design for achieving low model rate is in update sparsity. We achieve this by using a spike-and-slab prior around global model parameter values. Training with the spike-and-slab prior is sparsity enforcing. To quantify savings of our chosen spike-and-slab prior, we compute the cross-entropy of our prior with the resulting model δ after training. This is the number of bits we need to transmit the model update. On the other hand, assume we had a

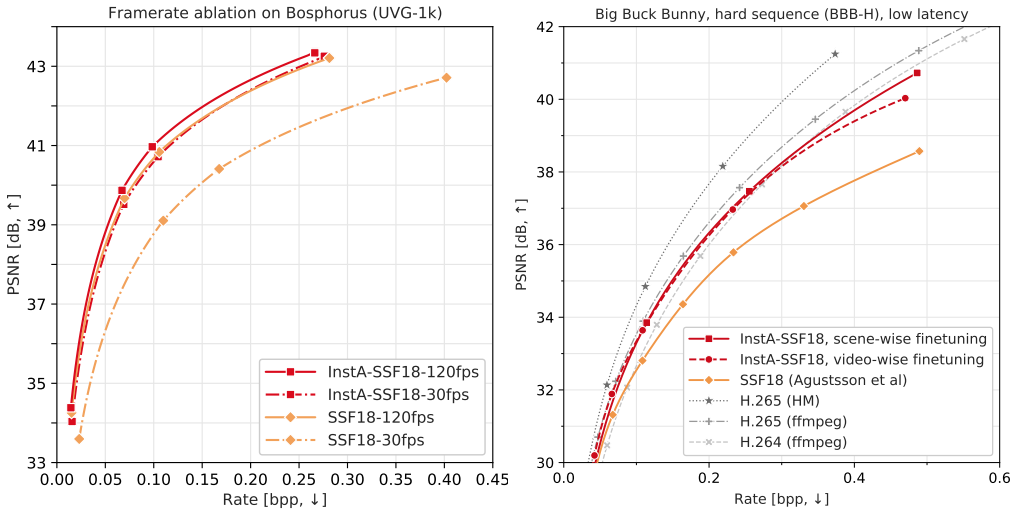

Figure 9: *Left:* Framerate ablation study. We compare the SSF18 baseline (orange) and InstA-SSF18 (red) on the original video at 120 fps (solid) and on the temporally subsampled video at 30 fps (dot-dashed). Instance-adaptive compression increases robustness to variation in framerate. *Right:* Full-video finetuning ablation study. We compare the performance of an SSF model finetuned on the full Big Buck Bunny video (10 minutes) to a SSF model finetuned only on the BBB-H scene (10 seconds) when evaluated on the BBB-H scene.

Gaussian prior with same variance as our "slab" component, and assuming the resulting δ follows the same distribution, the entropy of this Gaussian prior gives the expected number of bits to transmit non-sparse model update. We estimate the difference of the two scenarios to be around 7.5 bits per parameter.

Finetuning strategy for longer videos For longer videos with multiple scenes, one can consider several different strategies for instance-adaptive finetuning. The simplest approach is finetuning a single model on the full video. It may be beneficial to instead finetune the model independently on each scene. Even a hierarchical approach in which a model finetuned on the full video is further finetuned on each scene is an option.

The optimal strategy will in general depend on the particular video content and the similarity of the scenes as well as on the bitrate setting. We perform a simple first test by finetuning a single model on the entire Big Buck Bunny video, a 10 min video (14315 frames) with multiple scenes united by a common style. The results in the right panel of Fig. 9 show that such a single finetuned model leads to clear RD improvements over the global SSF model. However, in the large-bitrate region these improvements are smaller than the gains we achieve by finetuning on the BBB-E and BBB-H scenes individually.

H ETHICS STATEMENT

Video compression affects the life of millions of people and any algorithm and model should be carefully analyzed for potential ethical issues before deployment. In particular, it is important to check for harmful biases, which may be due to imbalanced training data, but can also be caused or exacerbated by algorithmic choices. To some extent, the instance-adaptive video compression method we are proposing may be helpful in combating such biases, as it can improve the compression performance on data that is different from typical samples in the training set. However, it cannot solve the problem entirely, as more pronounced instance adaptation will increase the bitrate.

The environmental impact of our research is also relevant. The energy cost of our experiments directly contributes to the climate catastrophe. In the long term, our work may also have a beneficial aspect, as it allows us to lower the bitrate and thus the energy cost of video streaming, which already now make up the majority of internet traffic.

I REPRODUCIBILITY STATEMENT

Our instance-adaptive video compression algorithm is described in detail in Sec. 3. The most complex components are the base compression models, which are described in more detail in the references we cite as well as in appendix B. With the details we provide, instance-adaptive video compression can be implemented and used for any base model with relatively little effort. Finally, in the supplementary material we provide the performance of the various methods on each video sequence as a CSV file to aid comparisons.