A neural video codec with spatial rate-distortion control

Noor Fathima, Jens Petersen, Guillaume Sautière, Auke Wiggers, Reza Pourreza
Qualcomm AI Research‡
{mohamedg, jpeterse, gsautie, auke, pourreza}@qti.qualcomm.com

Abstract

Neural video compression algorithms are nearly competitive with hand-crafted codecs in terms of rate-distortion performance and subjective quality. However, many neural codecs are inflexible black boxes, and give users little to no control over the reconstruction quality and bitrate. In this work, we present a flexible neural video codec that combines ideas from variable-bitrate codecs and region-of-interest-based coding. By conditioning our model on a global rate-distortion tradeoff parameter and a region-of-interest (ROI) mask, we obtain dynamic control over the per-frame bitrate and *the reconstruction quality in the ROI at test time. The resulting codec enables practical use cases such as coding under bitrate constraints with fixed ROI quality, while taking a negligible hit in performance compared to a fixed-rate model. We find that our codec performs best on sequences with complex motion, where we substantially outperform non-ROI codecs in the region of interest with Bjøntegaard-Delta rate savings exceeding* 60%.

1. Introduction

Lossy compression algorithms aim to compress given data in two ways: identifying redundancies, and selectively omitting parts of the data. These algorithms aim to find a tradeoff between the *bitrate*, *i.e.*, the number of bits spent to transmit the compressed representation, and *distortion*, the difference between the original data and its reconstruction.

Neural network-based video codecs, which learn to identify redundancies from example data, have seen great advances in rate-distortion (R-D) performance [23, 15, 1, 34, 42, 33]. Nevertheless, only few neural codecs can actually be deployed in a realistic video compression setting, as they are inflexible: typically, one model only supports coding at a particular bitrate *on average*, and the bitrate varies depending on the complexity of the data. Furthermore, most neural approaches do not distinguish semantically important regions of the video and background regions, which can be used to save bits with no perceptual quality degradation.

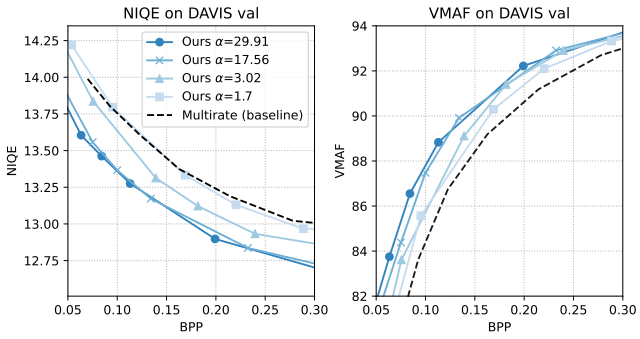

Figure 1: VMAF (↑) and NIQE (↓) on DAVIS val.

To overcome these issues, recent work aims to improve their flexibility. On one hand, variable-bitrate codecs enable coding under fixed-bitrate constraints [13, 25, 34] by controlling the per-frame bitrate using a control parameter. On the other hand, region-of-interest (ROI) based codecs [20, 8, 47, 31, 38] enable coding of semantically important regions with high precision by reallocating bits *within* the frame. In this work, we introduce a multi-rate ROI-based video codec that unifies these two concepts. The main challenges here are the lack of available ROI masks for video compression datasets, and the effect of temporal inconsistencies in ROI masks affecting reconstruction quality.

We address both issues by building on the artificial ROI mask generation procedure of Perugachi-Diaz *et al.* [31]. However, unlike [31], who train one model per tradeoff point, we train one model to cover multiple bitrate and ROI quality tradeoff points. The end result is a *multi-rate ROI-based* codec, with one parameter that controls the global rate, and another parameter that controls the relative distortion and bit allocation between ROI and non-ROI. Both can be adjusted dynamically at test time, leading to increased flexibility. We believe ours is the first video codec to offer this degree of test-time flexibility.

We evaluate our codec on DAVIS, a video dataset [32] with publicly available semantic annotations, and show that quality in the ROI is substantially improved compared to single-rate and multi-rate baselines with more than 60%

‡Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

Bjøntegaard-Delta rate (BD-rate) [5] savings in the ROI. Additionally, we extract ROI masks for standard video compression datasets to enable comparison to existing neural codecs, using HRNet [43] trained on MS-COCO, an off-the-shelf segmentation network. We observe the greatest success on videos with high motion content, such as videos of sports scenes and panning cameras, and see improvements in perceptual metrics such as VMAF [21] and NIQE [29], as shown in Fig. 1. However, applying our approach with a naive controller for ROI fidelity works less well for videos where the background is static, such as video conferencing. We provide evidence that this can be addressed using smarter rate-fidelity control algorithms, for example by improving the non-ROI quality of the first frame in the sequence.

In summary, the contributions of this work are:

(a) The first ROI-based multi-rate neural video codec that can dynamically control and adjust both local (within-frame) and global (per-frame) bitrate allocation.

(b) Qualitative and quantitative evaluation on common video datasets, demonstrating large gains in R-D performance for the region of interest, with BD-rate savings of up to 60%.

(c) Evidence that use cases like teleconferencing do not benefit from naive ROI coding, and that larger gains can be achieved in videos with high motion content, and by using smarter rate-fidelity control algorithms.

2. Background and related work

2.1. Neural data compression

Neural network-based data compression has been applied in many domains, including the image [41, 40, 3, 35, 28] and video [23, 36, 15, 14, 1, 34, 18] setting. In contrast to hand-engineered codecs such as HEVC [39], *neural codecs* learn to trade off bitrate and distortion from example data. The tradeoff is controlled by a hyperparameter β in the loss function:

$$\mathcal{L} = \mathbb{E}\left[\beta \mathcal{L}_{rate}(z) + \mathcal{L}_{dist}(x, \hat{x})\right]. \quad (1)$$

Here, z is a quantized latent variable obtained from an encoder, and $\hat{x}$ is the reconstruction obtained by passing z through a decoder network. Latent variables are further compressed using entropy coding by learning their distribution $p(z)$ using a context model or *prior* p_θ. Quantization of z is typically performed using quantization noise, differentiable relaxations, or a combination of both [3, 40].

At a given bitrate, the distortion loss $\mathcal{L}_{dist}$ with which the codec was trained determines what the resulting reconstructions look like. Mean squared error (MSE) is the common choice as it is directly related to the Peak Signal-to-Noise Ratio (PSNR) evaluation metric. Other well-established handcrafted metrics include Structural Similarity Index Measure (SSIM) [44] and its Multi-Scale variant (MS-SSIM) [45]. The rate loss $\mathcal{L}_{rate} = -\log p_\theta(z)$ is the negative log likelihood of the quantized latent under the prior, and is typically measured in bits per pixel.

2.2. Variable bitrate compression

Neural codecs are often trained for a single R-D tradeoff, *i.e.*, one model is trained with a pre-defined tradeoff parameter β. Using the loss from Eq. 1, the codec will reach the tradeoff point *on average*. However, video content with little redundancy or unpredictable movement will typically require more bits than the average. This poses a problem for practical settings where the maximum permissible bitrate is determined by a bandwidth, which cannot be exceeded for the more challenging frames. To support this use case, multiple single-rate models would have to be deployed, leading to memory overhead. Although solutions exist that allow training codecs to meet target bitrates exactly [37], a more common solution is to deploy a multi-rate codec, which supports a range of bitrates in a single model [9, 11, 13, 25, 31, 38, 25].

Some variable bitrate approaches change the quantization strategy through *latent scaling*. This method changes the quantization binwidth by scaling pre-quantization latents with a factor s before transmission [9, 13, 25, 31]. The main advantage of this approach is that, in theory, a single-rate codec can be turned into a variable-rate codec by only training a latent-scaling auxiliary network. A limitation is that the scaling factor s must be transmitted along with the latent z, but this cost typically amortizes over the size of the transmitted data.

Another common approach to training such models is to provide the R-D tradeoff parameter β as model input. During training, different β parameters are sampled from a pre-specified range, and the loss from Eq. 1 is changed accordingly. Examples of this approach can be found in the image [11, 38] and video [34, 49] settings. In this work, we utilize β-conditioning.

2.3. ROI-based compression

ROI-based compression, sometimes referred to as object-based coding or semantic compression, describes algorithms that are able to encode specified regions of interest with high fidelity. Unlike variable bitrate approaches that trade off rate and distortion at the "global" level, say on a per-frame basis, ROI-based methods allow controlling bitrate at a "local" level, for example within a frame. Neural ROI-based codecs typically achieve this by providing a spatial map as input to the model that indicates the region of interest [2, 8, 22, 38], although *latent scaling* has been

used for this purpose as well [25]. The loss is changed to incentivize the model to emphasize fidelity in the ROI, for instance by emphasizing the rate term for the non-ROI [47] or by de-emphasizing the distortion term the non-ROI [31].

Some neural image codecs extract the ROI as part of the compression algorithm. Cai *et al*. [8] and Li *et al*. [20] learn the ROI implicitly, masking latent variables with a spatial map generated by the encoder. Additionally, work for video exists that simultaneously extracts ROIs and compresses the data, by performing foreground-background separation inside the codec [46, 17]. Although this is an intuitive choice as foreground is likely to be important for perceptual quality, we opt to separate ROI extraction and compression in this work to maintain flexibility. This also enables optimizing the ROI mask for downstream tasks, as in [38].

Closest to our work is Perugachi-Diaz *et al*. [31], who introduce two ROI-based architectures based on the Scale-Space Flow (SSF) model [1]. We believe that details in (1) architecture and training scheme and (2) our evaluation set our work apart. For (1), while our base model is similar to the *implicit ROI SSF* of [31], we condition dynamically on α and β, whereas [31] train one model per α and β. Our changes lead to a much more practical codec, and enable a thorough investigation of global and local rate tradeoffs, in particular through the ROI vs non-ROI BD-rate plots in Fig. 3. Additionally, conditioning on these parameters enables rate control capabilities similar to standard codecs like H.265 [39] with a single model. For (2), we provide a more extensive evaluation than [31], including qualitative and quantitative analysis on standard video compression datasets UVG and HEVC-E2. The latter allows us to test the hypothesis of applicability of ROI-based coding to teleconferencing, one of the main motivations many works in the field [10, 16, 26]. We believe that our results show that *naive* ROI-coding may not be well-suited for static scenes like teleconferencing.

3. Method

3.1. Loss function

Similar to Perugachi-Diaz *et al*. [31], we use a loss that explicitly controls the global bitrate through a factor β, and controls fidelity in ROI versus non-ROI through a factor α and a given ROI mask m. This is equivalent to setting the following distortion loss (we omit parameters for the loss terms for brevity):

$$\mathcal{L} = \mathbb{E}\left[\beta\mathcal{L}_{rate} + \mathcal{L}_{dist}^{ROI} + \frac{1}{\alpha}\mathcal{L}_{dist}^{BG}\right] \tag{2}$$

$$\mathcal{L}_{dist}^{ROI}(x, \hat{x}, m) = m \odot (x - \hat{x})^2 \tag{3}$$

$$\mathcal{L}_{dist}^{BG}(x, \hat{x}, m) = (1 - m) \odot (x - \hat{x})^2 \tag{4}$$

Here, $\odot$ denotes the elementwise product and m contains binary values $\{0, 1\}$. High α corresponds to the setting where the background fidelity is less important, while $\alpha = 1$ corresponds to the case where foreground and background are equally important in expectation.

3.2. Architecture

The starting point for our model is the SSF model from Agustsson *et al*. [1], which we briefly explain here. SSF consists of an *I-frame* and a *P-frame* codec. The first frame in a sequence, x_0, is transmitted using the I-frame codec, and the remainder is transmitted by the P-frame codec. For each timestep $t > 1$, the motion between reconstruction $\hat{x}_{t-1}$ and the ground truth frame x_t is transmitted using the *P-frame flow model*, and an initial prediction $\hat{x}_t^{\text{warp}}$ is obtained using motion compensation. Then, a residual between this predicted frame and the target frame x_t is computed and transmitted using the *P-frame residual model*.

We modify SSF to enable global and local bitrate control. A visualization of our codec is shown in Fig. 2. For each frame, one global tradeoff parameter β_t, a local tradeoff parameter α_t, and a ROI mask m_t are input to the model. The ROI mask and α are combined to form a *weighted ROI mask* m_t^{α}. While α_t is not explicitly sent to the receiver, β_t is transmitted, as it is used to condition the decoders on the receiver-end. We dub the resulting architecture *Multi-Rate Multi-Distortion SSF (MR-MD-SSF)*. We emphasize that SSF no longer offers state-of-the-art performance, but it is a well-tested baseline that allows us to demonstrate the benefits of our multi-distortion approach. Importantly, there is an established open-source implementation [4], while more recent architectures like ELF-VC do not offer one. We initially tried to reproduce ELF-VC to use it as a base model, but did not manage to train it in a stable manner. Nevertheless, we believe our findings should translate to other architectures, as the conditioning blocks (see Fig. 2) can be added to any convolutional layer.

Global bitrate control: The model is conditioned on the rate tradeoff parameter β_t using conditional convolutions [11] that modulate feature maps using scaling and shifting parameters, as shown in Fig. 2 on the right. Specifically, for each activation in the network, a condition block outputs a per-channel scale and shift. We adopt the commonly used one-hot embedding representation for conditioning [11, 34], denoted here as β_t^*, where the range is discretized to a fixed number of bins. Instead of rounding β_t to the nearest matching bin, interpolation is used to obtain a soft embedding vector. Although there is no fundamental reason that a one-hot embedding should outperform simply providing β_t as a scalar input, we find that this results in easier training in practice. For all experiments, we use a one-hot embedding vector with 4 bins, and β_t is transmitted as a 16 bit integer.

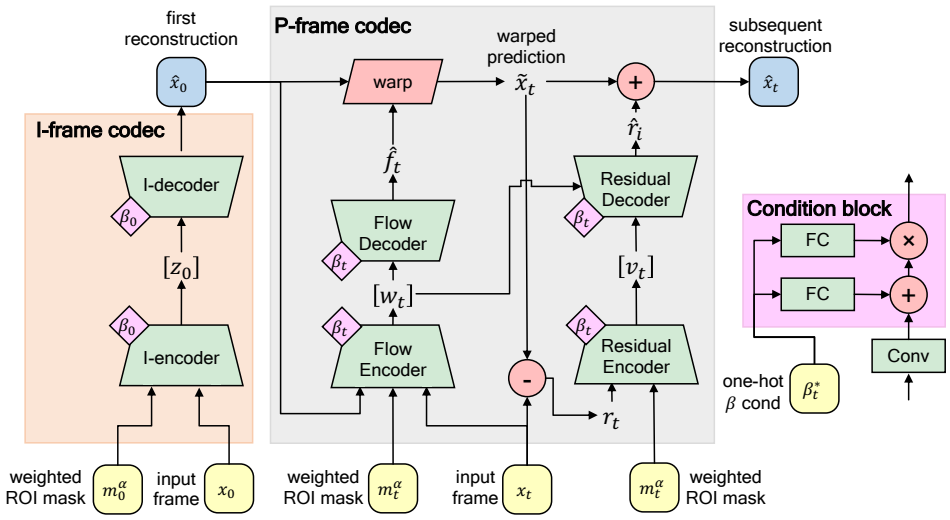

Figure 2: The MR-MD-SSF architecture. Activations in each encoder and decoder are scaled-and-shifted based on β_t, using the condition block shown on the right. β_t is transmitted to the receiver as a 16 bit integer.

Local bitrate control: The distortion tradeoff parameter α is provided to all frames in a video sequence. As α is used in Eq. 2 to trade off distortion in the ROI and non-ROI regions, it influences which information is sent in the compressed representation. Thereby, α indirectly influences bitrate allocation for the ROI and non-ROI regions. We combine α and the ROI mask m_t to create a weighted ROI mask for each timestep, m_t^α. This mask is provided as an input to each encoder in the I-frame and P-frame codecs. Early experiments showed that conditioning on α in a similar manner to β using conditional convolutional layers did not result in benefits over the current approach.

3.3. Training method

Sampling α and β factors: During training, the rate tradeoff parameter β is sampled from the range $[0.0001, 0.0128]$ for each video sequence in the batch. We sample β from a distribution skewed such that low β are sampled more often than high ones. The procedure is to first sample u_β from the skewed uniform distribution $U(0,1)^\gamma$, and use the mapping $u_\beta \rightarrow \beta$ as follows:

$$\log_2 \beta = (\log_2(\beta_{max}) - \log_2(\beta_{min})) \times u_\beta + \log_2(\beta_{min}) \quad (5)$$

We use $\gamma = 3$ for all experiments. After sampling β, it is embedded using the aforementioned embedding scheme to create the model conditioning β^*. This conditioning is used on both the encoder and receiver side.

The ROI tradeoff parameter α is sampled similarly to β: u_α is drawn from $U(0,1)$, and the conversion to α follows Eq. 5, but using $1 - u_\alpha$ as input so that the largest u_α corresponds to the lowest α. The latter is in the range $[1, 60]$, and one α is sampled for each video sequence. We combine the ROI mask m and the sampled u_α value to create the weighted ROI mask $m^\alpha = (1-m) \odot u_\alpha + m$, which is provided as input to the encoders. Note that we provide values u_α to the model, which are in the range $[0, 1]$, while the loss (Eq. 2) is weighted with α from the range $[1, 60]$. By providing the mask only as input to the encoder instead of using a separate model, the codec can learn when to transmit α and the ROI mask or when to omit them. The loss function ensures that the ROI mask will not be ignored.

For $\alpha = 1$, *i.e.* $u_\alpha = 1$, the conditioning mask m is 1 everywhere, which aligns exactly with setting where ROIs are not used. This design choice allows for easy comparison to non-ROI codecs, as we can compare evaluations with $\alpha = 1$ directly to baselines that do not take the ROI into account.

4. Experiments

4.1. Datasets

Unless specified otherwise, we use the Vimeo90K dataset [48] to train our model and SSF baselines. It contains 89,800 diverse sequences of scenes and actions. Following [31], we generate temporally consistent ROI masks use Perlin noise [30]. Note that these masks have no semantic connection to the underlying data.

For evaluation we use DAVIS [32] (val subset), a common benchmark for video segmentation, because of its high-quality annotations. As annotations are not available for compression benchmark datasets, we use an open-source implementation of HRNet [43] to predict masks for UVG [27] and HEVC class E2. More details on mask generation can be found in App. B.2.

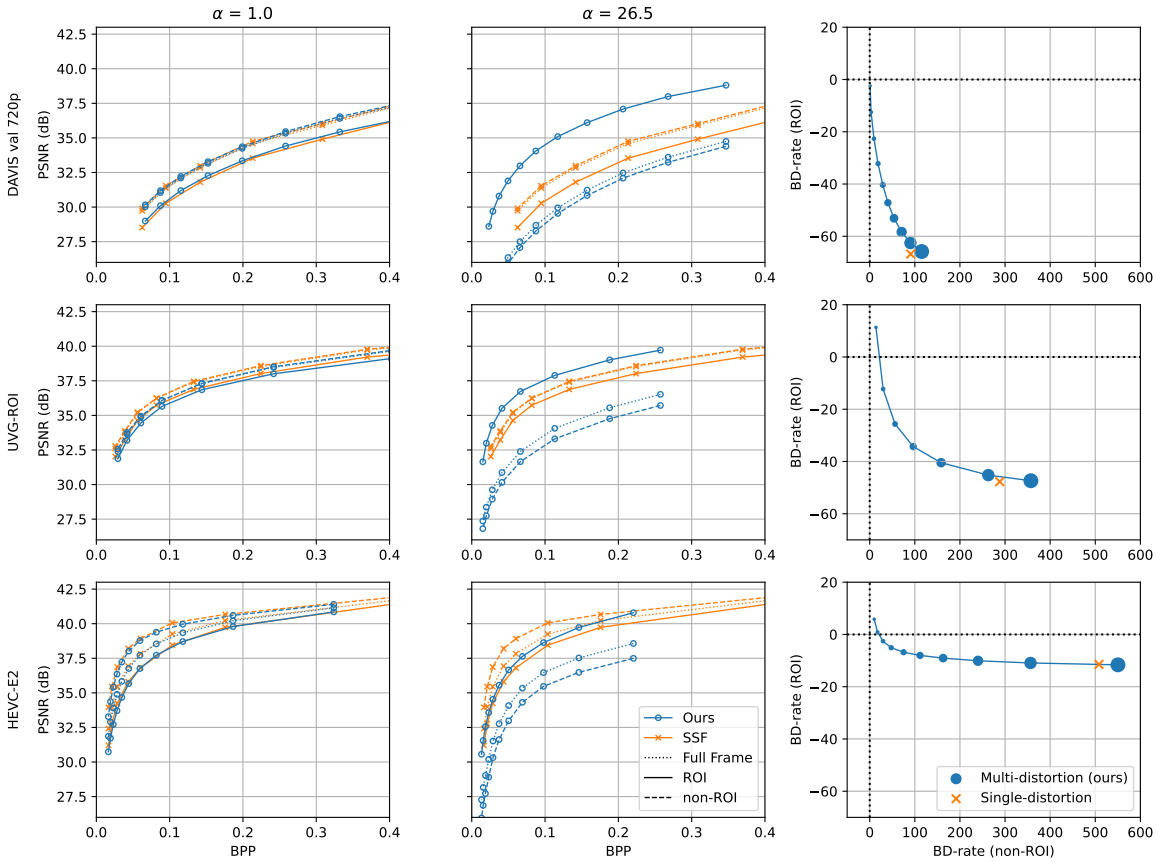

Figure 3: Left and center: R-D for $\alpha = 1$ and $\alpha = 26.5$. Right: ROI and non-ROI BD-rate with respect to SSF.

4.2. Training and evaluation details

Our model is warm-started from a SSF [1] model trained for single rate and distortion tradeoff ($\beta = 0.0001$, $\alpha = 1.0$) for 1M iterations, then finetuned for an additional 300k iterations using our architecture augmentations and ROI-aware loss $\mathcal{L}_{ROI}$. We crop video segments of 3 timesteps and size 256×256. We use the Adam optimizer [19] with a learning rate of 10^{-4} for the first 150k iterations, and reduce it to 10^{-5} for remaining iterations. For a fair comparison, all baseline SSF models are finetuned in the same way.

We compute per-frame scores in the RGB color space, then average over the frames of each video, and finally average over all videos to obtain a score for a given dataset. The results we report are based on group-of-picture (GOP) size of 12 for consistency with other neural compression works [24, 31, 33], *i.e.*, one I-frame is followed by 11 P-frames.

5. Results

5.1. Quantitative results

We summarize performance of our codec in Fig. 3. In the left and middle column, we show R-D performance for $\alpha = 1$ and $\alpha = 26.5$. Low α corresponds to the case where ROI and non-ROI are equally important, and we see that our model behaves similarly to the base SSF model that was trained using a single-rate and single-distortion objective. This is an important sanity check: it means that our training method and architecture augmentation enabled ROI-coding without important degradation to the regular coding performance. High α corresponds to the case where ROI PSNR is deemed more important than non-ROI PSNR. For the DAVIS and UVG datasets, ROI PSNR is substantially improved, at the cost of non-ROI PSNR. As the ROI typically only covers a small fraction of the frame for these datasets, the total PSNR depends mostly on non-ROI PSNR and drops as well.

To facilitate comparison across α-values, we plot the Bjøntegaard-Delta rate (BD-rate) [5] with respect to the SSF baseline in the rightmost column of Fig. 3. The BD-rate is a measure of distance between two R-D curves, summarized in a single scalar. In this visualization, ROI BD-rate and non-ROI BD-rate are plotted against each other to show how α influences performance across both axes. In these figures, each point on the curve corresponds to a u_α

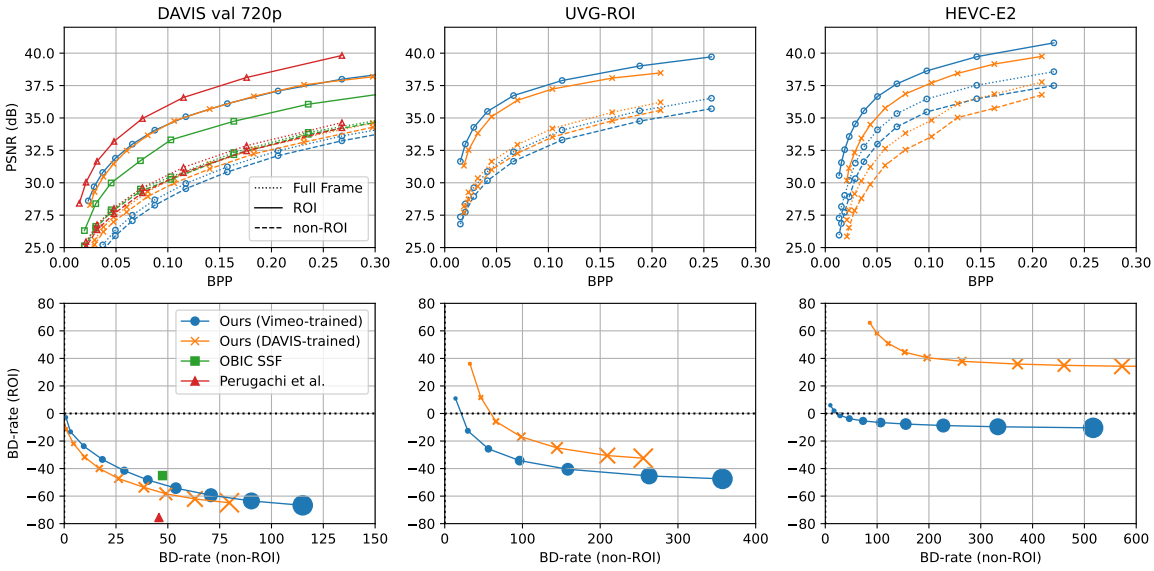

Figure 4: Vimeo (artificial masks) vs. DAVIS (real semantic masks) models. BD-rate x-axes are not shared.

value from the linear range $\{1.0, \ldots, 0.1\}$ (corresponding to $\alpha \in \{1.0, ..., 39.8\}$), and larger marker size indicates larger α (smaller u_α). To obtain one point, we compute BD-rate for the ROI and non ROI curve separately, with respect to the SSF baseline. The result is a curve that characterizes the tradeoff between ROI and non-ROI quality. To provide further intuition, $\langle x = 0, y = 0 \rangle$ means that performance is exactly equal to SSF in both ROI and non-ROI, $\langle x = +50, y = -50 \rangle$ means that BD-rate is increased by 50% in the non-ROI in order to save -50% BD-rate in the ROI. Positive BD-rate corresponds to worse performance.

We make the following observations. First, α controls the tradeoff between ROI and non-ROI for all datasets. Compared to an ROI-based model trained for a single α, shown as an orange cross, our model performs similarly, indicating that one model can navigate the tradeoff without a substantial hit in performance. Second, for low α, performance with respect to SSF is considerably better, up to 50% in the ROI and non-ROI. Third, performance improves further for higher α, until an inflection point is reached. Then, performance drops considerably. In summary, although α controls the tradeoff as expected, selecting it may require empirical evaluation.

Our ROI approach is more successful for natural videos in the DAVIS and UVG datasets than for the HEVC E2 dataset, which features content similar to video conferencing. The intuition for the limited gain in ROI quality for HEVC E2 is that even SSF automatically acts as a ROI codec when the background is static, as most of the bits will be spent on the dynamic foreground region, and coding of the background region is cheap. The main cause for poor non-ROI fidelity for the our ROI codec is the I-frame fidelity: the non-ROI is blurry when $\alpha = 26.5$, and subsequent P-frames inherit this low quality background. We analyze this further in Sec. 5.6, where we show that performance can be improved using better I-frame α-control.

Finally, we show perceptual distortion score VMAF [21] (↑) and no-reference quality metric NIQE [29] (↓) on the DAVIS-val dataset in Fig. 1. While the multirate baseline matches or outperforms our $\alpha = 1$ model, our models with higher α values obtain better scores over the entire range of bitrates. In other words, spending more bits on the ROI does increase overall perceptual quality, as captured by these metrics.

5.2. Comparison to literature

We first compare to the ROI *video* coding literature in Fig. 4. We show an SSF-based variant of the LearntOBIC ROI image codec [47], and "Implicit ROI SSF" video codec of Perugachi-Diaz et al. [31]. We take these numbers from [31], which means both baselines are trained on DAVIS-train. While we outperform OBIC, the Implicit ROI SSF beats our codec. As our architecture design is similar to the Implicit ROI SSF, we suspect the difference could be caused by dataset overfitting. This is further supported by our generalization study in Sec. 5.3, which shows that training on DAVIS-train leads to better overall performance on DAVIS-val.

Additionally, in Fig. 5, we compare the ROI-coding capability of our models with Song et al. [38], a ROI variable rate *image* codec. We evaluate our *MR-MD-SSF* in *all-intra* setting where every frame in a video sequence is encoded as

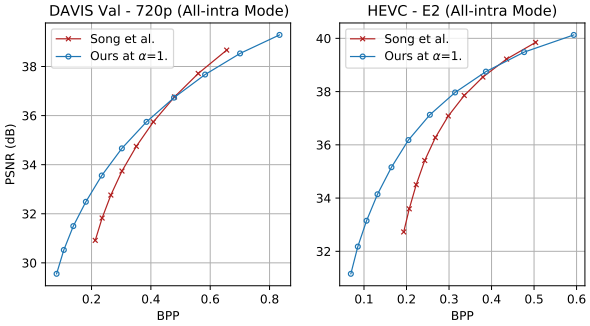

Figure 5: Rate-distortion performance of our model and Song et al. [38] in all-intra mode.

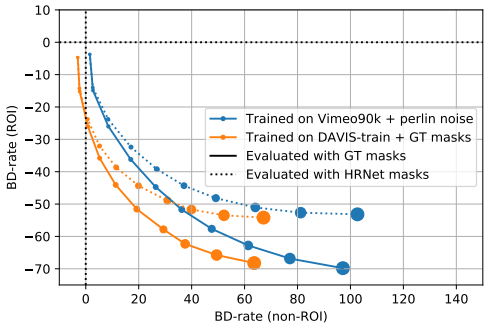

Figure 6: BD-rate curve on DAVIS-val. Solid line indicates ground-truth masks, dashed indicates predicted masks.

an *I-Frame*. We measure the R-D performance on DAVIS Val and HEVC-E2. Since the task-agnostic evaluation in [38] is done using a uniform quality mask with no distinction between ROI and non-ROI regions, we treat frames in a similar way by setting $\alpha = 1$. We observe better PSNR for our model at lower bitrates.

Extended comparisons to *non-ROI* codec literature can be found in App. D.1.

5.3. Training on synthetic ROI masks

Perugachi-Diaz *et al*. [31] report that training with artificial ROI masks works as well as training with real semantic annotations. We were not fully able to reproduce this finding, and refer the reader to App. B.1 for more details. Nevertheless, synthetic masks enable the use of larger unannotated datasets (such as Vimeo) for training, which should lead to better generalization.

In order to assess generalization performance, in Fig. 4, we show the performance for two versions of our model: the first is trained on the Vimeo dataset with artificial mask (blue), the second is trained on DAVIS with real masks (orange). We find that when evaluating on DAVIS-val, both models have similar performance, the DAVIS-trained one being slightly better than Vimeo trained model. However when evaluating the two models on UVG-ROI and HEVC-E2, the DAVIS-trained model exhibits a significant drop in performance, meaning the Vimeo-trained one generalizes far better. We hypothesize that the improved performance and generalization of the Vimeo-trained model is likely due to the larger dataset size, as training on DAVIS with artificial masks leads to a decrease in performance, which we discuss in App. B.1.

5.4. Effect of mask quality at test time

We use the official open-source pre-trained HRNet-OCR [43] trained on COCO-Stuff [7] to extract masks for the UVG [27] and HEVC-E2 [6] sequences (see App. B.2 for details). To get a sense of the quality of ROI masks generated with HRNet-OCR, we use it to extract masks for 10 DAVIS val sequences, and evaluate our model on these sequences using the predicted masks instead of the ground-truth annotations. We train one model on DAVIS with ground-truth annotations, and one model on Vimeo90k with synthetic masks. We evaluate both models on DAVIS-val, using either ground-truth masks, or the masks predicted using HRNet-OCR. We show BD-rate (with respect to SSF) in Fig. 6. For both models, we observe that for small α values, performance is similar for ground-truth masks and predicted masks. However, at high α, non-ROI BD-rate degrades quicker when using predicted masks. In conclusion, our model is moderately sensitive to "noisy" masks, especially for small α values, which is a likely operating point for most realistic use cases. Still, this result confirms that extracting high quality ROI masks at test time is crucial to produce reconstructions with high fidelity.

5.5. Qualitative results

In Fig. 7 we compare the SSF baseline to our model on the "Soapbox" sequence of DAVIS val. Each column pair compares SSF (left) to our model (right). The rightmost pair of columns is at half the rate of the one on the left. We show that although the bitrate is halved, our model (rightmost column) maintains PSNR in the ROI at around 33dB, while it drops significantly for SSF (second to last column), from 33 to 30dB. The difference is particularly salient if one focuses on the letters "S O" in the middle row close-up. Our model blurs the non-ROI region (bottom row), which is not as noticeable as it was already suffering from motion blur. We show additional qualitative examples in the App. C.

5.6. Error propagation in the non-ROI

We observe that teleconferencing videos rarely contain extreme scene changes, while background (i.e. note the human/face) occupies a significant part of the frame. Thus, in case of our MR+MD SSF model, always transmitting a low-quality background hurts performance, as errors in

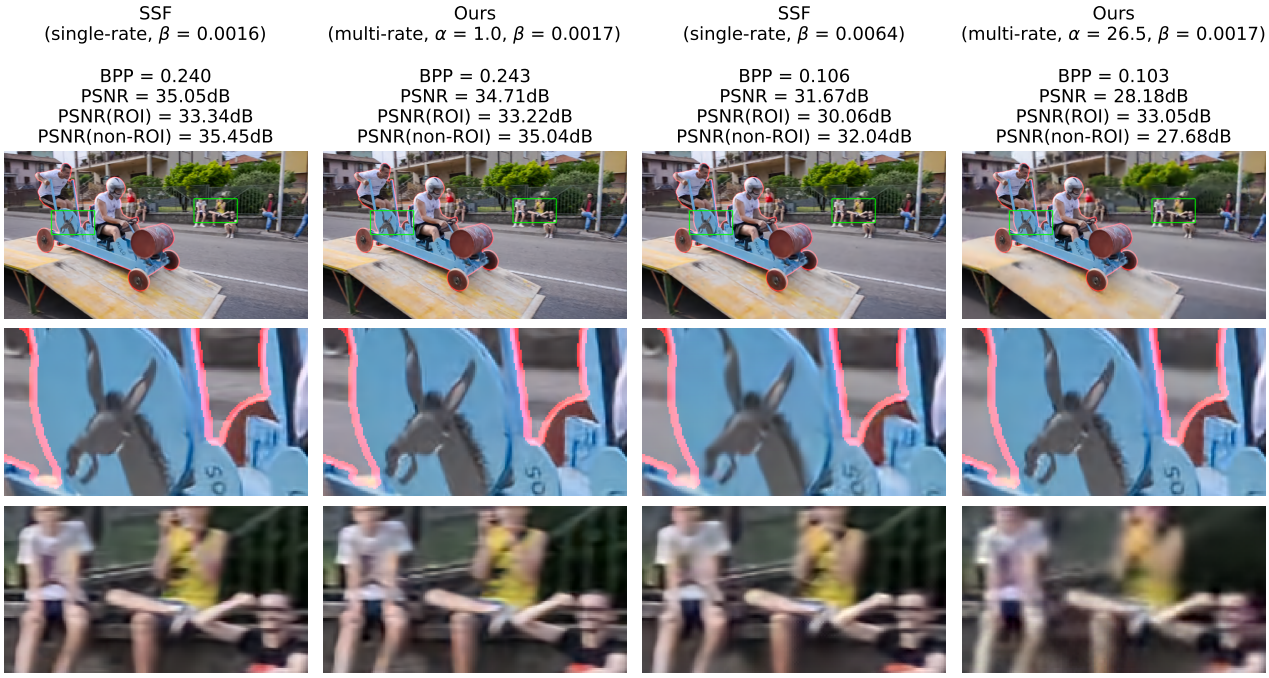

Figure 7: Close-up of reconstructions for frame 50 of the "Soapbox" sequence of DAVIS val, for single-rate, single-distortion SSF and our codec. The ROI mask is outlined in red.

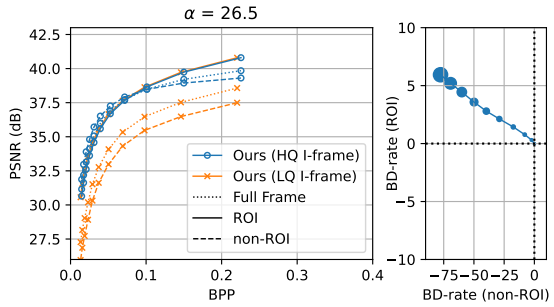

Figure 8: Left: R-D on HEVC E2. HQ I-frame codes I-frames with $\alpha = 1$, LQ I-frame uses $\alpha = 26.5$, all P-frames use $\alpha = 26.5$. Right: ROI vs. non-ROI BD-rate of HQ I-frame with respect to LQ I-frame.

I-frames propagate to P-frames. We propose a simple yet effective solution to alleviate this issue. We transmit a high-quality *I-frame* (at $\alpha = 1$) followed by *P-frame* at suitable α. This ensures that the background for the first frame is transmitted with the best possible quality, which improves non-ROI PSNR in subsequent P-frames. In Fig. 8 we show how effective the strategy is on HEVC-E2, where this simple change allows a boost of up to 80% BD-rate savings in the non-ROI, for a slight 5% BD-rate increase in the ROI. Similar analysis is provided for other datasets in App. D.4.

6. Conclusion

In this work, we introduce a variable-bitrate, region-of-interest based neural video codec. To the best of our knowledge, this is the first neural video codec that can dynamically adjust both the global (per frame) and local (within frame) allocation of bits. This is achieved by the introduction of two control parameters, one controlling the rate penalty as used in existing multi-rate models, and one controlling the tradeoff between distortion penalties in ROI and non-ROI regions. The resulting codec enables rate and quality control at a fine resolution, which has the potential to enable practical use cases such as coding at fixed rate with minimum ROI quality.

We demonstrate large BD-rate savings in the region of interest, achieving savings of above 60% in some cases. We also find that in the teleconferencing setting, the benefit of naive ROI-based coding is limited, and we provide intuition on how this issue can be mitigated.

Impact statement Learned codecs may be sensitive to bias in the data, which could lead to low quality reconstruction for data with little support. Additionally, semantic-aware codecs could be mis-used in surveillance settings. On the other hand, ROI-based neural codecs have the potential to improve fidelity in salient regions of videos important to the scene.

References

[1] 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 conference on Computer Vision and Pattern Recognition*, 2020.

[2] Hiroaki Akutsu and Takahiro Naruko. End-to-End learned ROI image compression. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[3] 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.

[4] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. CompressAI: a PyTorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020.

[5] Gisle Bjontegaard. Calculation of average psnr differences between rd-curves. *VCEG-M33*, 2001.

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

[7] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018.

[8] Chunlei Cai, Li Chen, Xiaoyun Zhang, and Zhiyong Gao. End-to-end optimized roi image compression. *IEEE Transactions on Image Processing*, 2020.

[9] Tong Chen and Zhan Ma. Variable bitrate image compression with quality scaling factors. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2163–2167, May 2020.

[10] Zhenzhong Chen, Junwei Han, and King Ngi Ngan. Dynamic bit allocation for multiple video object coding. *IEEE Transactions on Multimedia*, 2006.

[11] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Variable rate deep image compression with a conditional autoencoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3146–3154, 2019.

[12] Alex Clark. Pillow (pil fork) documentation, 2015.

[13] Ze Cui, Jing Wang, Bo Bai, Tiansheng Guo, and Yihui Feng. G-vae: A continuously variable rate deep image compression framework. *arXiv preprint arXiv:2003.02012*, 2020.

[14] Adam Goliński, Reza Pourreza, Yang Yang, Guillaume Sautière, and Taco S. Cohen. Feedback recurrent autoencoder for video compression. *ACCV*, 2020.

[15] Amirhossein Habibian, Ties van Rozendaal, Jakub M Tomczak, and Taco S Cohen. Video compression with rate-distortion autoencoders. In *IEEE International Conference on Computer Vision*, 2019.

[16] Sunhyoung Han and Nuno Vasconcelos. Object-based regions of interest for image compression. In *Data Compression Conference*, 2008.

[17] Trinh Man Hoang and Jinjia Zhou. RCLC: ROI-based joint conventional and learning video compression. *arXiv preprint arXiv:2107.06492*, 2021.

[18] Zhihao Hu, Guo Lu, and Dong Xu. FVC: A new framework towards deep video compression in feature space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1502–1511, 2021.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *iclr*, 2014.

[20] Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, and David Zhang. Learning convolutional networks for content-weighted image compression. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018.

[21] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 6(2), 2016.

[22] Jonas Löhdefink, Andreas Bär, Nico M Schmidt, Fabian Hüger, Peter Schlicht, and Tim Fingscheidt. Focussing learned image compression to semantic classes for V2X applications. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1641–1648, Oct. 2020.

[23] 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*, 2019.

[24] Guo Lu, Xiaoyun Zhang, Wanli Ouyang, Li Chen, Zhiyong Gao, and Dong Xu. An end-to-end learning framework for video compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3292–3308, 2021.

[25] Yadong Lu, Yinhao Zhu, Yang Yang, Amir Said, and Taco S Cohen. Progressive neural image compression with nested quantization and latent ordering. *arXiv preprint arXiv:2102.02913*, 2021.

[26] Marwa Meddeb, Marco Cagnazzo, and Béatrice Pesquet-Popescu. ROI-based rate control using tiles for an HEVC encoded video stream over a lossy network. 2015.

[27] 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*, MMSys '20, page 297–302, New York, NY, USA, 2020. Association for Computing Machinery.

[28] David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Neural Information Processing Systems*, 2018.

[29] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a completely blind image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[30] Ken Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3):287–296, 1985.

[31] Yura Perugachi-Diaz, Guillaume Sautière, Davide Abati, Yang Yang, Amirhossein Habibian, and Taco Cohen. Region-of-interest based neural video compression. *arXiv:2203.01978*, 2022.

[32] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv:1704.00675*, 2017.

[33] Reza Pourreza and Taco S Cohen. Extending neural p-frame codecs for b-frame coding. *IEEE International Conference on Computer Vision*, 2021.

[34] Oren Rippel, Alexander G Anderson, Kedar Tatwawadi, Sanjay Nair, Craig Lytle, and Lubomir Bourdev. Elf-vc: Efficient learned flexible-rate video coding. *Neural Information Processing Systems*, 2021.

[35] Oren Rippel and Lubomir Bourdev. Real-Time adaptive image compression. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2922–2930. PMLR, 2017.

[36] Oren Rippel, Sanjay Nair, Carissa Lew, Steve Branson, Alexander G. Anderson, and Lubomir Bourdev. Learned video compression. In *IEEE International Conference on Computer Vision*, October 2019.

[37] Ties van Rozendaal, Guillaume Sautière, and Taco S. Cohen. Lossy compression with distortion constrained optimization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 634–639, 2020.

[38] Myungseo Song, Jinyoung Choi, and Bohyung Han. Variable-rate deep image compression through spatially-adaptive feature transform. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2380–2389, 2021.

[39] 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.

[40] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. *International Conference on Learning Representations*, 2017.

[41] George Toderici, Damien Vincent, Nick Johnston, Sung Jin Hwang, David Minnen, Joel Shor, and Michele Covell. Full resolution image compression with recurrent neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5306–5314, 2017.

[42] Ties van Rozendaal, Johann Brehmer, Yunfan Zhang, Reza Pourreza, and Taco S Cohen. Instance-adaptive video compression: Improving neural codecs by training on the test set. *arXiv preprint arXiv:2111.10302*, 2021.

[43] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[45] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[46] Lirong Wu, Kejie Huang, Haibin Shen, and Lianli Gao. Foreground-background parallel compression with residual encoding for surveillance video. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[47] Qi Xia, Haojie Liu, and Zhan Ma. Object-based image coding: A learning-driven revisit. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.

[48] 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.

[49] Ruihan Yang, Yibo Yang, Joseph Marino, and Stephan Mandt. Insights from generative modeling for neural video compression. *arXiv preprint arXiv:2107.13136*, 2021.