Region-of-Interest Based Neural Video Compression

Yura Perugachi-Diaz*[1,2]
y.m.perugachidiaz@vu.nl
Guillaume Sautière[2]
gsautie@qti.qualcomm.com
Davide Abati[2]
dabati@qti.qualcomm.com
Yang Yang[3]
yyangy@qti.qualcomm.com
Amirhossein Habibian[2]
habibian@qti.qualcomm.com
Taco Cohen[2]
tacos@qti.qualcomm.com

[1] Vrije Universiteit Amsterdam
[2] Qualcomm AI Research, Qualcomm Technologies Netherlands B.V.
[3] Qualcomm AI Research, Qualcomm Technologies, Inc.

Abstract

Humans do not perceive all parts of a scene with the same resolution, but rather focus on few regions of interest (ROIs). Traditional Object-Based codecs take advantage of this biological intuition, and are capable of non-uniform allocation of bits in favor of salient regions, at the expense of increased distortion the remaining areas: such a strategy allows a boost in perceptual quality under low rate constraints. Recently, several neural codecs have been introduced for video compression, yet they operate uniformly over all spatial locations, lacking the capability of ROI-based processing. In this paper, we introduce two models for ROI-based neural video coding. First, we propose an implicit model that is fed with a binary ROI mask and it is trained by de-emphasizing the distortion of the background. Secondly, we design an explicit latent scaling method, that allows control over the quantization binwidth for different spatial regions of latent variables, conditioned on the ROI mask. By extensive experiments, we show that our methods outperform all our baselines in terms of Rate-Distortion performance in the ROI. Moreover, they can generalize to different datasets and ROI specifications at inference time. Finally, they do not require expensive pixel-level annotations during training, as synthetic ROI masks can be used with little to no degradation in performance. To the best of our knowledge, our proposals are the first solutions that integrate ROI-based capabilities into neural video compression models.

1 Introduction

The most common approach in neural lossy video compression is to rely on variational autoencoders to minimize the expected rate-distortion (R-D) objective, $D + \beta R$ [2, 17, 27,

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

*Work completed during internship at Qualcomm Technologies Netherlands B.V.

†Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc. and/or its subsidiaries.

31, 36]. Although this approach has proven to be successful, a model trained to minimize the expected rate-distortion tradeoff uniformly over all pixels may allocate too few bits to salient regions of a specific video. This clashes with the model of the human visual system, which is space-variant and has the highest spatial resolution at the the foveation point [21, 42]. Exploiting this phenomenon, *e.g.* by encoding Regions-Of-Interest (ROIs) with higher fidelity, can significantly contribute to the subjective quality under a low bitrate regime. The key idea of traditional *ROI-based codecs* [6, 10, 19, 24, 25, 37, 47] is to allocate different bitrate budgets for objects or regions of interest, and therefore to allow for non-uniform reconstruction qualities. For instance, traditional codecs like JPEG2000 [38] and MPEG-4 [41] were used as basis to build object-based coding methods [10, 19]. However, these ideas lacked widespread adoption due to their complexity and to their block-based nature, limiting their capability to deal with arbitrary ROI shapes.

More recently, some works have developed ROI-based neural *image* codecs, either by implicitly identifying the ROI as part of the encoding process [6, 24], or by relying on external algorithms for its extraction [47]. Under both approaches the R-D objective can be spatially weighted and, additionally, the latent variables can be masked before the quantization step to reduce their entropy [6, 17, 47]. Nevertheless, existing neural ROI-based codecs have the following limitations: *(i)* they only work for images, *(ii)* they use intricate masking schemes to spatially control the rate, without exploiting the Gaussian structure of the latent prior distribution and *(iii)* the encoding operations are tightly coupled with ROI prediction, which makes it hard for the codecs to be adapted to different ROI requirements.

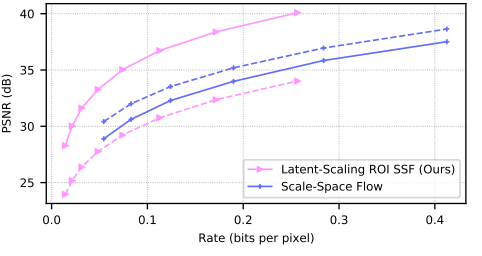


Figure 1: R-D improvements on DAVIS [33], where ROI-PSNR (solid) is higher than non-ROI PSNR (dashed). The improvement is equivalent to 69.3% BD-rate gain [4].

In this paper, we present the first two neural codecs capable of ROI-based compression. The first *implicit* model is fed with the ROI mask and is trained with an ROI-aware loss, where the distortion of the background is de-emphasized. Secondly, the *latent-scaling* model extends the implicit model by exploiting a recent technique originally developed for variable rate coding [8, 9, 13, 29]. We extend its design by introducing an auxiliary autoencoder (AE) being fed with the ROI map, and regressing a gain tensor explicitly controlling the quantization binwidth for different spatial regions. This can be seen as the continuous equivalent of the masking scheme used in conjunction with scalar quantization [6, 24, 30]. We describe our solution in the context of a Scale Space Flow (SSF) [2] architecture; however, we argue that they are in principle compatible with most state-of-the-art models based on hyperpriors [27, 34, 36]

We show that our methods outperform all our baselines on the DAVIS dataset [33] in terms of R-D performance, as measured in PSNR in the ROI (Fig. 1). Moreover, further analyses show that they generalize to any arbitrary ROI which can be specified by the user at inference time and that expensive pixel-dense annotations are not required during training, as synthetic ROI can be used with little to no degradation in performance.

2 Related work

Non-uniform coding. The literature on spatially variant image encoding mainly focuses on two separate problems: *(i)* how to estimate the ROI and *(ii)* how to exploit it to improve coding. Most traditional block-based methods [10, 18, 19] fall under the former category, and simply exploit non-uniform coding capabilities of standard codecs such as JPEG2000 [38] and MPEG-4 [41]. These solutions are limited in their capabilities due to their block-based approach to compression, which hinders the encoding of arbitrarily shaped objects and does not allow for pixel-level bit allocation optimisation [25, 37].

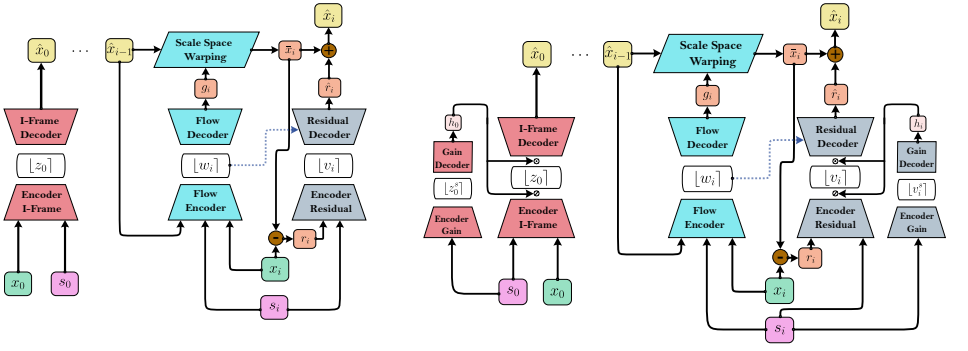
In contrast, recent work in neural image coding tackle both the above mention problems and target pixel-level ROI [1, 3, 6, 14, 15, 24, 47]. Among these, Li *et al*. [24] and Cai *et al*. [6] learn the ROI implicitly by spatially masking out the latents before scalar quantization, whereas Xia *et al*. [47] use the down-scaled output of the DeepLab [7] segmentation network to mask out foreground from background, before sending each stream to a separate hyper-codec for quantization. Similar to these works, our work focuses on how to use a given ROI to enable non-uniform coding, whilst delegating its extraction to some external automatic model such as [7, 22, 23, 43, 44, 45, 50]. However, our approach extends extends neural ROI-coding to the case of video inputs.

Neural video compression. Compressing videos with neural networks has been an active field of research recently [2, 16, 17, 20, 27, 28, 34, 35, 36, 46]. While varying in their choice of architecture and quantization strategy, neural video codecs generally follow the DVC [28] framework where an I-frame codec compresses the first frame and a P-frame codec uses motion estimation and a residual network to model the subsequent ones. Recently, Agustsson *et al*. [2] proposed to use a Scale-Space Flow which addresses uncertainties in motion estimation via interpolation through a Gaussian pyramid. This allows blurring of the warped frames in regions where optical flow prediction is uncertain or ill-posed, like chaotic motions and obstructed objects. Our work is established in the same SSF framework, and enables ROI-based coding by means of latent scaling [8, 9, 13, 29], a technique originally introduced for variable bitrate coding. Differently from these works, that scale the latents globally with a single scalar value, we adjust the quantization step size for every spatial location, thus controlling the levels of distortion and entropy in foreground and background regions.

In summary, we are the first work to learn ROI coding end-to-end for video inputs (as opposed to images) and extend latent-scaling spatially to be used in an ROI-based context. Additionally, other works either learn implicitly the ROI using a subnetwork [6, 15, 24] or tie themselves to a restricted set of semantic classes [1, 3, 14], which would require re-training if testing on unseen classes. In contrast, we explicitly take the ROI as input, which provides the user evaluation time flexibility similar to H.264 and H.265 ROI mode.

3 ROI-based neural video compression

In this section we first present the neural video codec we use as backbone for our work, Scale Space Flow (SSF) [2]. Next, we extend SSF to be an ROI-based codec by proposing two models: the *Implicit* and *Latent-scaling* ROI SSF. Lastly, we will describe the optimization for SSF and the ROI-aware methods.



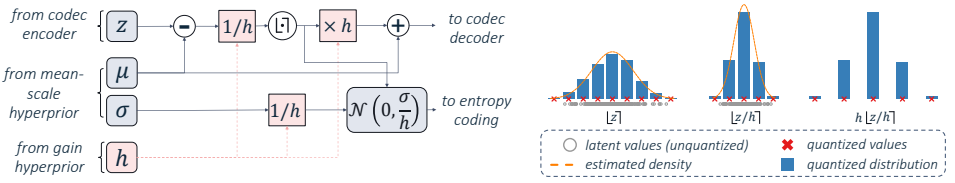
(a) Implicit ROI Scale-Space Flow (b) Latent-scaling ROI Scale-Space Flow

Figure 2: Illustration of the proposed ROI-based neural video compression models. Both models learn to utilize an ROI mask s. Model (a) feeds a mask along with an image and model (b) utilizes extra hyperpriors to send the ROI mask for latent scaling.

We define a video frame $x_i \in \mathbb{R}^{H\times W\times 3}$ at time step i, where H and W represent its height and width respectively. Then, a video sequence is denoted as $\mathbf{x} = \{x_0, x_1, \ldots, x_T\}$, with $T+1$ frames. The sequence of binary ROI masks corresponding to the video sequence is defined as $\mathbf{s} = \{s_0, s_1, \ldots, s_T\}$, where $s_i \in \{0,1\}^{H\times W}$. The neural video codec SSF consists of an I-frame codec and a P-frame codec. The I-frame codec is a mean-scale hyperprior AE [31] which encodes a first frame x_0 independently to produce a reconstruction $\hat{x}_0$. The P-frame codec is comprised of two hyperprior AEs. The first, the *P-frame flow hyperprior AE*, estimates a scale-space flow g_i from the previous reconstruction $\hat{x}_{i-1}$ and current frame x_i, which is used to warp the previous reconstruction into $\bar{x}_i$. The second hyperprior AE, *the P-frame residual hyperprior AE*, encodes the residual $r_i = \bar{x}_i - x_i$. The final reconstruction $\hat{x}_i$ is obtained by adding the warped prediction $\bar{x}_i$ and the estimated residual $\hat{r}_i$. The latent codes of each hyperprior AE are denoted by z_0, w_i and v_i and are rounded to integer values then entropy coded using the prior parameters estimated by their respective hyper-decoder. We omit hyper latent codes for ease of exposition, and we refer to [2] for further details.

Implicit ROI Scale-Space Flow An immediate extension to SSF to make it ROI-aware is to provide the ROI mask s_i as input to each of the three hyperpriors, see Fig. 2a. Note that the ROI mask is not fed to the decoder, meaning we expect the encoders to implicitly store the relevant ROI information inside the existing latent codes. Since the decoder does not require the ROI mask, we do not need to transmit a representation of the mask itself. Feeding information of the mask along with the video frame, in combination with the use of an ROI-aware loss, encourages the model to focus on important aspects for the user. Albeit simple, we show the effectiveness of this approach when paired with an ROI-aware loss in Sec. 4.

Latent-scaling ROI Scale-Space Flow Inspired by methods like [5, 24] which introduce a mechanism to explicitly control the spatial bit allocation, we adapted a recent technique called latent-scaling [8, 13]. Albeit similar in its motivation, it differs from the masking approach of [13] by exploiting the Gaussian prior structure of mean-scale hyperprior AE. The key idea is to apply a scaling factor to the latent which changes the quantization step size, leading to different trade-offs between rate and distortion in ROI and non-ROI areas. By using ROI-based information to control the scale of latents, the quantization grid can



(a) (b)

Figure 3: Illustration of the latent scaling mechanism for ROI-based coding. (a) shows how latent-scaling variable h affects latent z and prior parameters μ and σ. (b) shows intuition on why scaling the prior is necessary for entropy coding.

be explicitly adjusted. Our model can therefore learn that foreground regions require finer quantization than background regions. For ease of exposition, we will describe in the next paragraphs latent-scaling for the I-frame hyperprior AE, but the same method is applied to the P-frame residual hyperprior AE. We do not apply it to the P-frame flow hyperprior AE as initial studies showed the flow code w_i only accounts for a small fraction of the total rate. For similar reasons, we only apply latent-scaling latents, leaving hyper-latents, which are cheap to encode, unaffected.

We introduce a new hyperprior-like network called *gain hyperprior AE* (see leftmost autoencoder in Fig. 2b). This network encodes the ROI mask s_0 into a latent code z_0^s, that is decoded to a gain variable h_0 which shares the same dimensions as the latent variable z_0, both spatially and channel-wise[‡]. We scale the latent z_0 with the inverse of the estimated spatial gain variable h_0, where we restrict $h_0 \geq 1$. Such a procedure is akin to making the quantization range larger, depending on the value of h_0. We further denote the mean μ and scale σ as the prior parameters estimated by the I-frame hyper-decoder. In the quantization step, we choose to center the scaled latent $z_0 \oslash h_0$ by its prior mean $\mu \oslash h_0$, where $\oslash$ is a elementwise division. Next, we apply the rounding operator $\lfloor \cdot \rceil$ on $(z_0 - \mu) \oslash h_0$ such that the estimated mean μ learned by the hyper-encoder is on the grid, and then add the offset $\mu \oslash h_0$ back. The dequantized latent $\hat{z}_0(h_0)$ is obtained by multiplying by h_0 after the quantization block. More precisely:

$$\hat{z}_0(h_0) = \lfloor (z_0 - \mu) \oslash h_0 \rceil \odot h_0 + \mu, \tag{1}$$

where $\odot$ denotes elementwise multiplication. After the dequantized latent $\hat{z}_0(h_0)$ is obtained, it is passed to the decoder to obtain reconstructed frame $\hat{x}_0$. The whole procedure is illustrated in Fig. 3a. For rate computation and entropy coding, we use the modified probability $\mathbb{P}$ of $\hat{z}_0(h_0)$ as follows:

$$\mathbb{P}\left(\hat{z}_0(h_0)\right) = \int_{\hat{z}_0(h_0)-h_0/2}^{\hat{z}_0(h_0)+h_0/2} \mathcal{N}(x-\mu|0,\sigma)dx \tag{2}$$

$$= \int_{\hat{z}_0(h_0)/h_0-1/2}^{\hat{z}_0(h_0)/h_0+1/2} \mathcal{N}\left(x - \frac{\mu}{h_0}\middle|0, \frac{\sigma}{h_0}\right) dx \tag{3}$$

As shown in Fig. 3b and in Eq. (2), latent-scaling can be interpreted as effectively changing the quantization grid / binwidth. In practice, for entropy coding we do not change the quantization grid and round to the integer grid and scale the prior appropriately, as in Fig. 3a and b (middle plot) and Eq. (3). As stated above, the same procedure is applied to the P-frame residual latent code v_t, as shown in Fig. 2b.

[‡]previous latent-scaling [8, 13, 25] work only use channel-wise gain

ROI-aware Rate-Distortion Loss We modify the regular R-D loss from SSF to take into account the ROI mask. We sum the rate and distortion for all T frames in the video sequence $\mathbf{x}$ with corresponding ROI masks $\mathbf{s}$:

$$\mathcal{L} = \beta \mathcal{L}_R + \sum_{i=0}^{T} \mathcal{L}_{D,i}, \tag{4}$$

where β is rate-distortion trade-off variable. $\mathcal{L}_D$ represents the distortion loss which is a modified mean squared error (MSE) involving the binary ROI mask:

$$\mathcal{L}_{D,i} = \frac{1}{HWC} \sum_{j=1}^{H} \sum_{k=1}^{W} \sum_{l=1}^{C} \left(s_i \odot \varepsilon_i + \frac{1}{\gamma} \cdot (1 - s_i) \odot \varepsilon_i \right)_{jkl}, \tag{5}$$

where H, W and C denote the image dimensions, γ is a penalty hyperparameter for the non-ROI, $\varepsilon_i = (x_i - \hat{x}_i)^2$ is the squared error and s_i is broadcasted over the channel dimension. Note that the distortion loss of the original SSF corresponds to the special case where s_i equals one everywhere. Further, the rate loss $\mathcal{L}_R$ is computed with the estimated cross-entropy $\mathcal{H}(\cdot)$ by the hyperprior of each latent variable present in the model. For the implicit ROI SSF the rate loss $\mathcal{L}_{I,R}$ is equal to:

$$\mathcal{L}_{I,R} = \mathcal{H}(z_0) + \sum_{i=1}^{T} \left[\mathcal{H}(v_i) + \mathcal{H}(w_i) \right]. \tag{6}$$

The rate loss $\mathcal{L}_{LS,R}$ of the latent-scaling ROI SSF also includes latent variables z_0^s for the latent scaling of the I-frame hyperprior AE and v_i^s for the latent scaling of the P-frame residual hyperprior AE. As such, it is given by:

$$\begin{aligned} \mathcal{L}_{LS,R} =& \mathcal{H}(z_0^s) + \mathcal{H}(z_0) \\ &+ \sum_{i=1}^{T} \left[\mathcal{H}(v_i^s) + \mathcal{H}(v_i) + \mathcal{H}(w_i) \right]. \end{aligned} \tag{7}$$

In practice we found that the two extra rate contributions from the ROI masks $\mathcal{H}(z_0^s)$ and $\mathcal{H}(v_i^s)$ are only a small fraction compared to the standard rate components $\mathcal{H}(z_0)$ and $\mathcal{H}(v_i)$ of the model. Please note that in both Eq. 6 and 7 we omit the rate of the hyper latent codes to avoid notational clutter.

4 Experiments

Datasets. As standard video compression benchmarks [5, 40, 48] do not come with ROI annotations, we hereby introduce a benchmark for ROI-based codecs, by utilizing publicly available video segmentation datasets and deriving ROI maps from their pixel-level groundtruth labels. More specifically, we rely on DAVIS [33] and Cityscapes [12] for training and evaluation of our models. DAVIS is composed of 90 diverse and short video sequences, for which groundtruth segmentation of salient objects provided. To create binary ROI masks, we consider all labeled objects as foreground, whereas the rest of the frame is labeled as background. We use 60 sequences for training and 30 for validation, comprising 4,209 and 1,999 frames respectively. Cityscapes is composed of 2,120 video sequences from dashcam of vehicles driving around German cities. 1,885 sequences are used for training and

235 for validation, or 89,248 and 15,000 frames respectively. As groundtruth segmentation labels are provided only at 1 fps, we extract semantic labels automatically for every frame by running the state of the art segmentation model in [39]. The dataset provides a categorization of every pixel into one of 19 classes. We select pixels of "*vehicle*", "*road*", "*pedestrian*", "*bicycle*", "*motorcycle*" as belonging to the ROI, and mark other classes as non-ROI. To reduce compression artifacts, we resize the frames from both datasets to 720p using Pillow [11].

As an alternative to ground-truth ROI masks, in some experiments (see Sec. 4)) we rely on synthetic ROI masks generated using Perlin noise [32] (only during training). The masks contain blobs that evolve continuously over time to cover each of the video frames.

Implementation details. We optimize all methods but SSF with the ROI-aware MSE as distortion metric (Eq. (5)), and use $\gamma = 30$ as penalty for the non-ROI areas. Following the training scheme from [2, 34], all models are warm-started from an SSF pre-trained on the Vimeo-90k dataset [49] for 1M steps, then fine-tuned on the dataset of interest for 300K steps. We trained all models at various rate-distortion tradeoffs with $\beta = 2^{\alpha} \times 10^{-4} : \alpha \in 0, 1, ..., 7$. We use Adam optimizer with a learning rate of 10^{-4} with batch size 8. Each example in the batch is comprised of 3 frames (I-P-P), randomly cropped to 256×384. The models take about 3 days to train on a single NVIDIA V100 GPU. We report video quality in terms of PSNR in ROI and non-ROI, where both are first calculated per-frame in the RGB color space, then averaged over all the frames of each video, and finally averaged over all the videos of a dataset. The results we report are based on Group-of-Picture of size of 12 for consistency with other neural compression works [2, 27, 28, 34]. We refer to the appendix for architecture details, along with information about the computational complexity of the models.

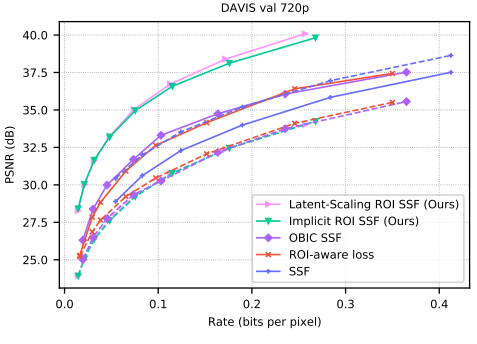


Figure 4: All ROI-based neural video compression approaches vs SSF. Solid line denotes ROI PSNR, while dashed non-ROI PSNR.

Compared methods. We compare our method to the plain SSF and two further ROI-based baselines. The first, dubbed *ROI-aware loss*, consists of SSF trained with our ROI-aware loss as described in Eq. (4). While the codec is blind to the ROI, it is expected to implicitly learn it through the training objective, in a similar fashion as the semantic models in Habibian *et al.* [17]. The second method, dubbed *OBIC SSF*, is based on a recent ROI-based neural image codec [47]. To enable a fair comparison, we train this architecture using our ROI-aware loss, which is slightly different from the formulation in [47].

ROI-based coding In Fig. 4, we report the RD-plots of Implicit ROI SSF and Latent-scaling ROI SSF. We compare our proposed models to the described ROI-aware loss and OBIC SSF baselines, as well as to a plain SSF model that does not involve any ROI-based compression.

For all compared models, solid lines and dashed lines correspond to RD curves in ROI and non-ROI regions respectively. The figure shows several insights. First, the plain SSF shows better compression results on non-ROI regions, that are seemingly easier to compress than ROI areas on DAVIS. This result - that we hypothesize is due to the high degree of motion affecting foreground objects on the dataset - underlines that such a codec might be suboptimal. The ROI-based baselines we consider, namely ROI-aware loss and OBIC SSF, succeed in delivering a better tradeoff for foreground regions. Overall, their performances seem comparable across the rate spectrum. Interestingly, the separate hyperprior models envisioned by OBIC SSF for foreground and background barely outperforms a simple ROI-aware loss in our experiments. Finally, the figure clearly shows the superiority of the proposed implicit and latent-scaling ROI SSF. Indeed, their RD-curves performs on par with the mentioned baselines on background regions, while achieving a superior tradeoff for ROI regions. In this respect, our latent-scaling based model seems to slightly outperform the implicit model in ROI areas, especially at higher bpps (> 0.1).

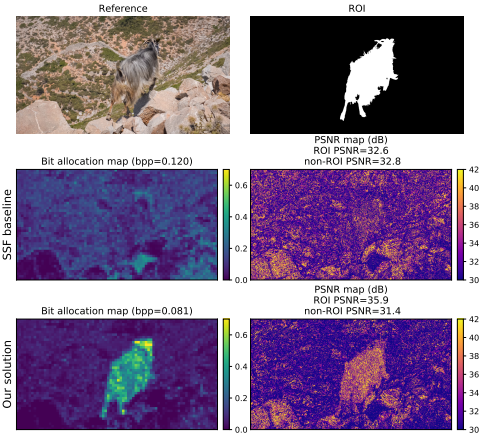


Figure 5: Bitrate and PSNR allocation maps for SSF and our proposed ROI-based codec, latent-scaling ROI SSF. We hereby report frame 5 of DAVIS "goat" sequence.

Furthermore, we investigate the behavior of the proposed Latent Scaling ROI SSF codec in terms of spatial bit allocation and reconstruction quality. Fig. 5 shows, on a reference validation frame from DAVIS, the pixel-wise bpp and PSNR as compared to the ones achieved by SSF. For SSF, bit allocation and reconstruction quality are roughly uniformly distributed over the image. Differently, Latent Scaling ROI SSF model focuses both bpp and PSNR on the region of interest. Moreover, it is worth noting how, despite the fact latent scaling operates at the reduced resolution of the latents (resulting in block-wise bpp allocation), the PSNR of the reconstructed frame properly aligns with the ROI at pixel-level. Finally, in Fig. 6 we shows a few qualitative compression results of our model, compared to SSF.

Generalization We investigate the generalization capability of our proposed latent-scaling ROI SSF model to different data and regions of interest. To do so, train a model on DAVIS and measure its performance on Cityscapes. We expect (at least) two main sources of generalization gap. First, the videos in the two datasets depict very different content (*data gap*), and differences in the acquisition settings may generate discrepancies in low-level image statistics and global motion[§]. Moreover, the ROI specification described above might impact training (*ROI gap*). To monitor both effects, we plot in Fig. 7a the RD curves of our latent scaling model and plain SSF, trained either on DAVIS or on Cityscapes, and evaluated on Cityscapes. By considering the gap between the SSF model (blue lines) trained on DAVIS and the one trained on Cityscapes, we notice how the former performs slightly worse than

[§] for instance, in Cityscapes the motion is dominated by the ego-motion of the camera, which is car-mounted.

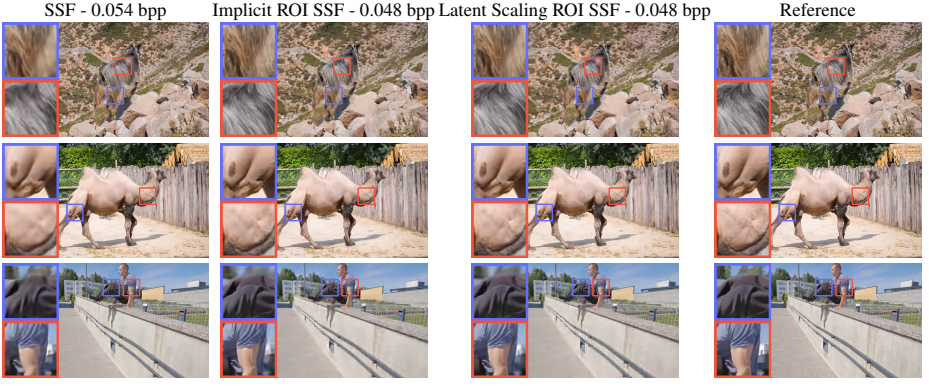
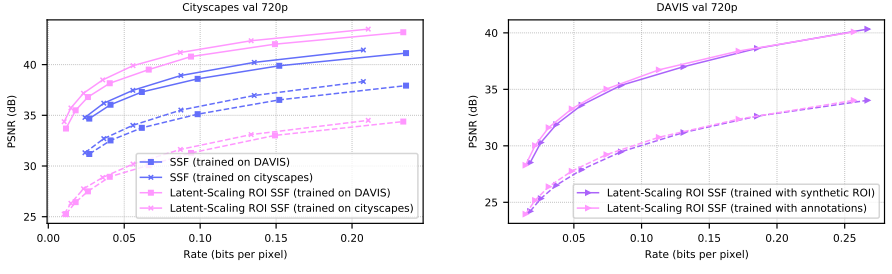


Figure 6: Qualitative results of SSF and our proposed ROI-based codecs, implicit ROI SSF and latent-scaling ROI SSF, on the sequences "`goat`, `camel`, `parkour`" sequence of DAVIS Val 2017. We hereby report frames 5, 11, 31 respectively.



(a) (b)

Figure 7: (a) Latent-scaling ROI SSF tested on Cityscapes. (b) Effect of training with synthetic ROI instead of ground-truth annotation for the binary ROI mask. Solid / dashed lines denote ROI/non-ROI PSNR respectively.

the latter, both for ROI and non-ROI areas. This gives a sense of the severity of the data gap alone, as no ROI was employed during training whatsoever. In order to assess the effect of the ROI gap, we examine the margin between the two trainings of Latent Scaling ROI SSF (pink lines). Interestingly, we observe a similar edge as the one observed for plain SSF. The fact that the performance gap does not increase significantly suggests that most of the discrepancy is still due to the data gap, and that our codec is barely susceptible to the nature of ROIs used during training. Finally, we observe that, when evaluated on Cityscapes ROI areas, the ROI-based model trained on DAVIS outperforms the SSF model. This observation suggests that, when interested in ROI-based compression on a target dataset, our codec trained on a different dataset might still be a better choice than its non ROI-based counterpart, even when the latter is trained on the target dataset itself.

Synthetic ROI masks In order to further investigate the sensitivity of our latent scaling based codec to the nature of ROIs used during training, we carry out an experiment where we train it using synthetically generated masks. Specifically, we rely on the DAVIS dataset and we generate the ROI for every training clip randomly, by taking advantage of perlin noise [32]. The resulting masks are temporally smooth, but do not correlate with the content of the video itself. In Fig. 7b we plot the performance of such a model (in purple) against a

model trained on regular semantic masks, obtained by manual annotation (in pink). We emphasize that both models are tested, on the validation set, on regular semantic masks of ROI objects. Thus, we expect the model trained on realistic ROI masks to trace an upper bound RD-curve for the model trained on synthetic. Interestingly, results show that a gap exists between the two models, but it is almost negligible, confirming the intuition that our model is minimally affected by the nature of training ROIs. The close performance represented in RD-curves suggests that, although in the case ROI masks are available at training time their use is worthwhile, their lack does not represent a serious impediment for optimizing the model, as the use of synthetic masks yields similar performance on realistic use cases.

5 Limitations and societal impact

Our main motivation for the Latent Scaling ROI Scale-Space Flow was to allow for inference-time single model multirate behavior for the largest rate model, without the need to re-train or to adapt the training scheme like in [13, 36] (similar to what was demonstrated in [29] for image compression). This would make our ROI codec more practical to deploy by drastically reducing the number of parameters and allowing fine-grain control of the rate. However, it does not allow for a fully multirate model (*i.e.* a single model covering the whole rate spectrum), and it comes with an increase in implementation complexity with minor performance benefits over the simpler implicit ROI approach.

In addition, visual assessments highlighted how, in their current implementation, both ROI-based models can sometimes produce sharp quality transitions between ROI and non-ROI regions. The problem would probably be exacerbated if the ROI masks suffered both in terms of quality and in temporal consistency. Both of these issues may be overcome by using smooth masks during training and/or inference.

Finally, a user study would benefit the evaluation of quality of the compressed videos as quantitative quality metrics were shown to poorly correlate with human judgment [26]. Such an analysis, based on subjective metrics such as Mean Opinion Scores (MOS), would further confirm that higher fidelity in the ROI at the cost of fidelity in the non-ROI can lead to a net boost in perceptual quality.

Concerning societal impact, we do not see immediate harmful applications of our method that might negatively affect any public. Note that because the ROI codecs depend on an ROI retrieval algorithm, the methods may suffer from (and potentially amplify) its biases and shortcomings.

6 Conclusions

In this paper, we introduced two methods for ROI-based neural video compression, capable of allocating more bits to pre-specified regions of interest in order to increase their fidelity. More specifically, we introduced an implicit model being fed with the ROI, as well as a latent scaling model explicitly controlling the quantization bitwidth of the latent variables in a spatial variant fashion. Both models are optimized by a ROI-aware rate-distortion objective. We showed that our methods outperform all baselines in terms of Rate-Distortion performance in the regions of interest, and that they can generalize to different datasets at inference time. Finally, they do not require expensive pixel-level annotations during training, as synthetic ROI masks can be used with little to no degradation in performance.

References

[1] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2019.

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

[3] Mohammad Akbari, Jie Liang, and Jingning Han. Dsslic: Deep semantic segmentation-based layered image compression. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

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

[5] Frank Bossen. Common test conditions and software reference configurations. JCTVC-F900, 2011.

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

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[8] Tong Chen and Zhan Ma. Variable bitrate image compression with quality scaling factors. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[9] Tong Chen, Haojie Liu, Zhan Ma, Qiu Shen, Xun Cao, and Yao Wang. Neural image compression via non-local attention optimization and improved context modeling. *arXiv preprint arXiv:1910.06244*, 2019.

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

[11] Alex Clark. Pillow (pil fork) documentation, 2015. URL https://buildmedia.readthedocs.org/media/pdf/pillow/latest/pillow.pdf.

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016.

[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] Shiyu Duan, Huaijin Chen, and Jinwei Gu. Jpad-se: High-level semantics for joint perception-accuracy-distortion enhancement in image compression. *arXiv preprint arXiv:2005.12810*, 2020.

[15] Yiping Duan, Yaqiang Zhang, Xiaoming Tao, Chaoyi Han, Mai Xu, Cheng Yang, and Jianhua Lu. Content-aware deep perceptual image compression. *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2019.

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

[17] 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.

[18] Sunhyoung Han and Nuno Vasconcelos. Image compression using object-based regions of interest. In *IEEE International Conference on Image Processing*, 2006.

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

[20] Zhihao Hu, Guo Lu, Jinyang Guo, Shan Liu, Wei Jiang, and Dong Xu. Coarse-to-fine deep video coding with hyperprior-guided mode prediction. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2022.

[21] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, 2004.

[22] Qiuxia Lai, Wenguan Wang, Hanqiu Sun, and Jianbing Shen. Video saliency prediction using spatiotemporal residual attentive networks. *IEEE Trans. on Image Processing*, 2019.

[23] Trung-Nghia Le and Akihiro Sugimoto. Video salient object detection using spatiotemporal deep features. *IEEE Transactions on Image Processing*, 2018.

[24] 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.

[25] Shipeng Li and Weiping Li. Shape-adaptive discrete wavelet transforms for arbitrarily shaped visual object coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 2000.

[26] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward A Practical Perceptual Video Quality Metric. *Netflix Technology Blog*, 2016.

[27] Jianping Lin, Dong Liu, Houqiang Li, and Feng Wu. M-lvc: Multiple frames prediction for learned video compression. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2020.

[28] 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.

[29] 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.

[30] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Conditional probability models for deep image compression. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018.

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

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

[33] 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.

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

[35] 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.

[36] 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.

[37] T. Sikora and B. Makai. Shape-adaptive dct for generic coding of video. *IEEE Transactions on Circuits and Systems for Video Technology*, 1995.

[38] A. Skodras, C. Christopoulos, and T. Ebrahimi. The jpeg 2000 still image compression standard. *IEEE Signal Processing Magazine*, 2001.

[39] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.

[40] Ultra Video Group. UVG test sequences. http://ultravideo.cs.tut.fi/, 2020. Accessed: 2020-02-21.

[41] A. Vetro, Huifang Sun, and Yao Wang. Mpeg-4 rate control for multiple video objects. *IEEE Transactions on Circuits and Systems for Video Technology*, 1999.

[42] Brian A. Wandell. Foundations of vision ma sunderland, 1995.

[43] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, and et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. doi: 10.1109/tpami.2020.2983686.

[44] W. Wang, J. Shen, J. Xie, M. Cheng, H. Ling, and A. Borji. Revisiting video saliency prediction in the deep learning era. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[45] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[46] Chao-Yuan Wu, Nayan Singhal, and Philipp Krähenbühl. Video compression through image interpolation. *Proceedings of the European Conference on Computer Vision*, 2018.

[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] Xiph.org. Xiph.org video test media [derf's collection]. https://media.xiph.org/video/derf/. Accessed: 2020-02-21.

[49] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T. Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 2019.

[50] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017. doi: 10.1109/cvpr.2017.660.