Spatial Rate Allocation for Learning-based Video Coding

Mohsen Abdoli, Félix Henry, Gordon Clare
Institute of Research and Technology b-com
Cesson-Sévigné, France
firstname.lastname@b-com.com

Abderrahmane Jarmouni, Kra-Tchimbie Koffi
UFR ISTIC - University of Rennes
Rennes, France
firstname.lastname@etudiant.univ-rennes.fr

***Abstract*—This paper presents a method that enables arbitrary end-to-end Learning-based image/video codecs to apply spatial rate allocation. At the frame-level, the forward pass of the underlying encoder network is followed by a latent refinement step, in which a customized loss function is minimized. This loss function takes as input an arbitrary pixel-wise map that defines the interest of each pixel and computes a weighted distortion with respect to the given interest map. Back-propagation of the customized loss function using the gradient descent gives a refined version of the frame latent in which the quality of regions of interest (ROI) is improved at the cost of quality of regions of disinterest. The proposed method is implemented on top of an existing end-to-end LVC, called AIVC[1], using salience-based interest maps. Experiments show that the proposed method can effectively improve the quality of regions of interest frames. Notably, BD-BR performance using Weighted PSNR (WPSNR) shows an improvement of up to 21% by the proposed method.**

***Index Terms*—Video coding, Neural networks, Rate allocation**

I. Introduction

Conventional video compression has been challenged by Learning-based Video Coding (LVC) in the past few years [1]–[5]. Despite challenges to address, this technology is rapidly maturing, offering an alternative perspective to the digital video communication systems. Several advantages could possibly be delivered by such learning-based system, compared to existing ones, such as High Efficiency Video Coding (HEVC) and Versatile Video Coding (VVC). First of all, the end-to-end nature of the optimization in learning-based systems allows a more global tuning of codec and avoid local and hand-made parameter tuning, as is the case with current codecs. Moreover, hardware implementation of an LVC system can possibly be agnostic to the underlying hardware. This is advantageous since updating a decoder would be as simple as updating the network model weights, while in conventional codecs, same update would require expensive chipset replacement.

LVC systems have currently several practical difficulties to tackle. More known objectives in this regard are namely: inexpensive hardware support especially at the receiver side to enable real-time decoding, standardization in different levels such as network abstraction, transport *etc.*, and last but not least, inflexibility at the encoding time by a learning-based encoder. The problem that this paper attempts at tackling is to enable spatial rate allocation at the encoding time. This problem is important to address since state-of-the-art LVC designs offer an encoding which is an unbending forward-pass of the encoder network. This is in contrast with conventional encoders, where arbitrary parameter exploration by the Rate-Distortion Optimization (RDO) process is allowed at the encoding time and user is free to choose an RDO strategy that best fits its constraints and objectives.

The problem of encoding-time spatial rate allocation by LVC has been sparsely studied in the literature [6]–[10]. RDOnet imitates the RDO process of the conventional codecs in a generic manner which might possibly allow spatial rate allocation. RDOnet deploys masking layers to zero-out certain coefficients. By training models with such layers, unimportant regions of the image are identified during inference and do not have their information transmitted [11]–[13]. In [14], an ROI-based multi-rate codec is proposed that can dynamically control local and global rate allocation at the frame-level. Moreover, PLONQ [15] presents a latent scaling based variable bitrate solution that defines multiple quantization levels with nested quantization grids and progressively coding of all latents across different quantization levels.

This paper proposes an alternative solution by using a frame-level latent fine-tuning that can virtually apply any spatial rate allocation strategy in order to better preserve quality of certain frame areas. This method is orthogonal to the underlying encoder and can be used on any pre-trained learning-based codec. The rest of this paper is organized as follows. Section II formalizes a base encoder on top of which the proposed method, described in Section III, is implemented. Section IV presents the experiment results of the implementations and finally, Section V concludes the paper.

II. Base coder

A pre-trained codec pair is given in the form of an encoder and decoder models, respectively expressed as $<\mathbf{E}, \mathbf{D}>$, whose parameters $\Theta =< \Theta_e, \Theta_d >$ are jointly optimized. Using this Learned Video Codec (LVC), encoding a frame x transforms it to a latent representation z in a latent space, as:

$$z = \mathbf{E}(x; \Theta_e). \tag{1}$$

Subsequently, decoding the latent representation z transforms it back to a pixel-domain representation $\hat{x}$:

[1] https://github.com/Orange-OpenSource/AIVC

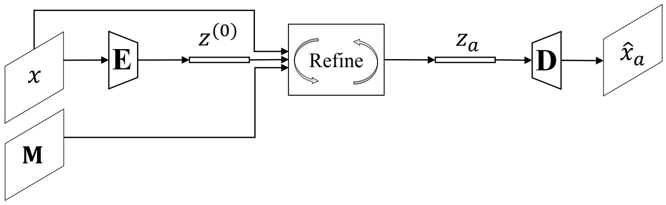

Fig. 1: A high-level view of the proposed frame-level latent refinement for rate allocation. Modules $\mathbf{E}$ and $\mathbf{D}$ denote performing the forward pass of the encoder and decoder networks, respectively. z_a is the output latent of this process whose decoding results is $\hat{x}_a$ which better preserves regions of interest.

$$\hat{x} = \mathbf{D}(z; \Theta_d). \tag{2}$$

The above round-trip is typically monitored by the two metrics of rate and distortion. Since the entropy coding operates exclusively on discrete data, thus is non-differentiable. Consequently, during the training process of an LVC, the rate is typically estimated by an approximated Probability Distribution Function (PDF) of the continuous uniform distribution p, as:

$$\tilde{\mathbf{r}}(z) = -\text{log}_2 p_z(z). \tag{3}$$

Moreover, the distortion metric $\mathbf{d}$ is computed as the squared l^2-norm of the compression loss error:

$$\mathbf{d}(x, z) = \|x - \hat{x}\|_2^2 = \|x - \mathbf{D}(z; \Theta_d)\|_2^2. \tag{4}$$

Given a Lagrangian multiplier λ, the training of the codec is performed on a dataset of samples $\mathbf{x} = \{x_k | k = 0, 1, ..., |\mathbf{x}| - 1\}$, by minimizing an estimation of its rate-distortion cost as the loss function:

$$\mathcal{L}(\mathbf{x}; \Theta) = \frac{1}{|\mathbf{x}|} \sum_{x_k \in \mathbf{x}} \mathbf{d}(x_k, z_k) + \lambda \tilde{\mathbf{r}}(z_k). \tag{5}$$

The optimization of the model parameters Θ to minimize the loss function of Eq. 5 is typically carried out by iteratively applying gradient descent with back-propagation of its error, expressed as $\nabla_\Theta \mathcal{L}(\Theta)$ in Eq. 6. Each iteration j of this algorithm back-propagates the error with a given learning rate parameter, denoted as η.

$$\Theta^{(j+1)} = \Theta^{(j)} - \eta \nabla_\Theta \mathcal{L}(\mathbf{x}; \Theta^{(j)}). \tag{6}$$

In the proposed rate allocation method of this paper it is assumed that an operational codec pair is provided as $< \mathbf{E}, \mathbf{D} >$, whose optimal parameters Θ^* are optimized by using Eq. 6 on a dataset.

III. Rate allocation with latent refinement

A. Interest map and weighted distortion

It is assumed that regions of interest in each image are provided as input and the way they are computed is out of the scope the proposal of this paper. Therefore, a pixel-wise map called the interest map is given for each image to code, which quantifies the relevance of pixels and is interpreted as the priority of preserving fidelity of each pixel.

Depending on the underlying problem, the interest map is computed differently. One common strategy is to choose regions of interest by taking into account the subjective relevance of objects/regions of the image. Alternatively, in video communication application, one might use the temporal motion flow as an indicator of pixels' objective relevance, to identify pixels that are more likely to be used as reference in motion compensation of next video frames. Either way, the goal is to allocate more rate to regions of interest in order to limit their decoded distortion.

Let $\mathbf{M}_{W \times H} = \{m_{ij} \mid i\text{=1,2,..,}W,\ j\text{=1,2,..,}H\}$ be the interest map of current image of resolution $W \times H$, where $0 \leq m_{ij} \leq 1$ quantifies the interest in quality preservation of pixel at position (i, j). A weighted distortion metric $\mathbf{d}^\omega$ is defined in Eq. 7 that takes into account the interest map $\mathbf{M}$ using the element-wise product (noted as $\odot$).

$$\mathbf{d}^\omega(x, \hat{x}, \mathbf{M}) = \|M \odot (x - \hat{x})\|_2^2 = \sum_{j=1}^{H} \sum_{i=1}^{W} (m_{ij}(x_{ij} - \hat{x}_{ij}))^2, \tag{7}$$

B. Latent refinement

Fig. 1 shows a high-level view of the proposed spatial rate allocation, given a interest map $\mathbf{M}$. This algorithm is implemented as a latent refinement process that is carried out after the forward pass of the encoder network. Given an initial latent of image x, calculated as $z^{(0)} = \mathbf{E}(x; \Theta_e^*)$, and its associated interest map $\mathbf{M}$, the latent refinement process gives a new latent representation, expressed as:

$$z_a = \text{Refine}(z^{(0)}, \mathbf{M}). \tag{8}$$

Inside the Refine operation, an iterative process progressively traverses the initial latent $z^{(0)}$ through N intermediate latents $z^{(t)}$ to eventually reach a latent z_a, where the desired rate defined by the interest map $\mathbf{M}$ is allocated. At each iteration t, a weighted RD cost is calculated on $z^{(t)}$ by using Eq. 3 and Eq. 7 for the rate and distortion computation, respectively. As the desired rate allocation scheme is already incorporated at the pixel-level in Eq. 7, using the same Lagrangian multiplier λ will result in the desired weighted rate-distortion trade-off. The weighted RD cost is expressed as:

$$\mathcal{L}^\omega(z^{(t)}, \mathbf{M}; \Theta^*) = \mathbf{d}^\omega(x, z^{(t), \mathbf{M}}) + \lambda \tilde{\mathbf{r}}(z^{(t)}). \tag{9}$$

Note that there are two main differences between the refinement loss $\mathcal{L}^\omega$ (Eq. 9) and the training loss $\mathcal{L}$ (Eq. 5). First, $\mathcal{L}^\omega$ is computed on a single sample, while $\mathcal{L}$ is computed

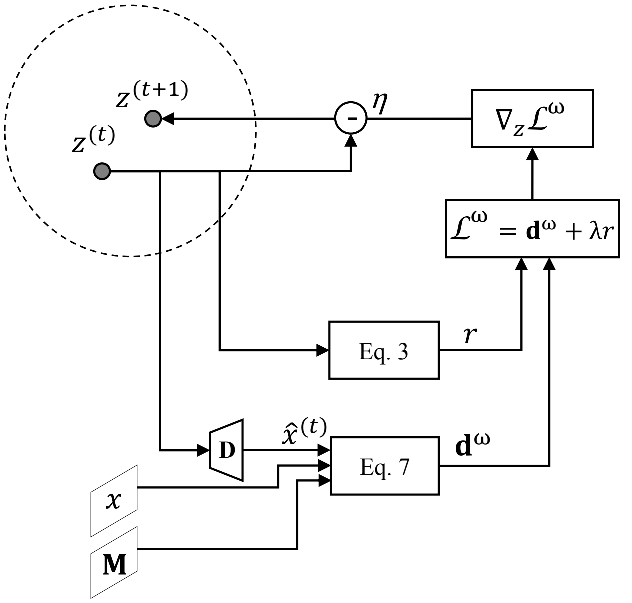
$z^{(t+1)}$
$z^{(t)}$
-
η
$\nabla_z \mathcal{L}^{\omega}$
$\mathcal{L}^{\omega} = \mathbf{d}^{\omega} + \lambda r$
Eq. 3
r
D
$\hat{x}^{(t)}$
Eq. 7
$\mathbf{d}^{\omega}$
x
M

Fig. 2: One iteration of the latent refinement, taking the input latent $z^{(t)}$ to the refined latent $z^{(t+1)}$.

on a dataset $\mathbf{x}$. Second, the uniform distortion $\mathbf{d}$ is replaced by $\mathbf{d}^{\omega}$ which applies the interest map $\mathbf{M}$.

The gradient of the refinement loss with respect to the latent z is expressed as $\nabla_z \mathcal{L}^{\omega}$. In contrast to $\nabla_{\Theta} \mathcal{L}$ (Eq. 6) that is computed with respect to model parameters Θ during the training phase, the gradient used in the refinement is computed with respect to the latent of the input signal. As a result, back-propagation of this error updates only the latent and keeps Θ unchanged:

$$z^{(t+1)} = z^{(t)} - \eta \nabla_z \mathcal{L}^{\omega}(z^{(t)}, \mathbf{M}; \Theta^*), \tag{10}$$

where η is the learning rate decay for an improved convergence and computed with two parameters of initial learning rate $\eta^{(0)}$ and decay rate β:

$$\eta^{(t)} = \eta^{(0)}/(1 + \beta.t) \tag{11}$$

Fig. 2 visualizes one iteration of the above process. This iteration starts with $z^{(t)}$ and ends with $z^{(t+1)}$, which could then be used as the input to the next iteration. Once adequate number of iterations of the above process are applied on the initial latent $z^{(0)}$, the refined latent which applies the given rate allocation strategy is produced as z_a.

IV. Experiments

A. Anchor

The "base coder" with no spatial rate allocation strategy is the main anchor of the experiments in this section. To this end, an open-source LVC called AIVC which has shown a competitive performance in the CLIC challenge of the past few years, has been used [16].

One might accurately argue that latent refinement even without rate allocation (i.e. with uniform rate allocation) most likely improve the performance as well. Thus, as the benchmark, the base coder integrated with latent refinement using a uniform interest map, precisely with a constant interest as $m_{ij} = C$ for all values of i and j within the image. This is also the equivalent of using Eq. 4 instead of Eq. 7, for distortion computation of Eq. 9. This benchmark is called "base refinement" in the rest of this section. Furthermore, only All-Intra configuration is used in all experiments presented in this paper.

B. Metrics

The main performance metric used in this paper is Bjontegaard-Delta Bit-Rate (BD-BR) [17]. However, as spatial rate allocation disrupts the usual trade-off of rate and quality, it is expected that the overall performance of the codec – which is computed unbeknownst of regions of interest – will decrease. Therefore, to quantify the performance of the proposed spatial rate allocation, the importance of regions is incorporated in the quality metric to weight pixels according to their interest WPSNR [18].

C. Interest map generation

A binary map $\mathbf{B}$ is first generated to identify the region(s)-of-interest. To do so, a salience prediction algorithm presented in [19] is used. Then, this binary map is translated into the actual interest map $\mathbf{M}$ by:

$$m_{ij} = \begin{cases} C_{roi}, & \text{if } \mathbf{B}(ij) = \text{true} \\ 1 - C_{roi}, & \text{if } \mathbf{B}(ij) = \text{false}, \end{cases} \tag{12}$$

where $0.5 < C_{roi} < 1$ determines how much we prioritize the quality preservation of the region(s)-of-interest during the latent refinement iterations. In the experiments of this paper, $C_{roi} = 0.6$ has been chosen empirically. Moreover, in order to guarantee that the spatial rate allocation would be aligned with the predefined relative importance of the distortion with respect to the rate, the interest map $\mathbf{M}$ is normalized such that $\sum_{j=1}^{H} \sum_{i=1}^{W} m_{i,j} = W \times H$.

As mentioned earlier, the performance and technical aspects of the salience prediction algorithm is not important as long as it outputs a reasonably relevant interest map. Fig. 3 shows examples of the salience maps generated by the adopted method on some of the test sequences of this paper.

D. Performance

Table I presents the BD-BR performance of the proposed method as well as the "base refinement" benchmark, both against the "base coder" anchor. Since both refinement methods are iterative, results are provided for different number of iterations, namely one, five and twenty iterations. As can be seen, the proposed method improves the base coder as well as the base refine in all settings. Precisely, the proposed latent refinement for rate allocation method improves the base encoder by -3.6%, -11.4% and -21.2%, while the based latent refinement improves the base encoder by -2.7%, -6.6% and -11.5% for settings of 1-iter, 5-iter and 20-iter, respectively.

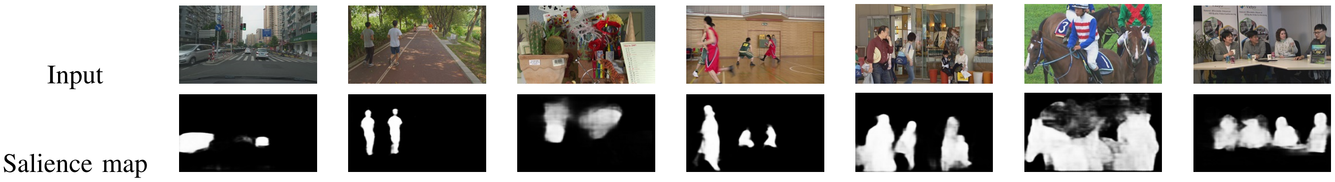

Fig. 3: Examples of the adopted salient zone prediction applied on the first frame of the test sequences.

TABLE I: BD-BR performance (with WPSNR) of the proposed rate latent refinement with salience-based rate allocation (Proposed refine) versus the base refinement with no rate allocation.

Class	1-iter		5-iter		20-iter	
	Base refine	Proposed refine	Base refine	Proposed refine	Base refine	Proposed refine
A1	-2.3%	-3.8%	-6.7%	-10.3%	-12.1%	-21.1%
A2	-3.0%	-3.9%	-6.6%	-11.5%	-11.6%	-21.3%
B	-3.0%	-4.0%	-7.2%	-11.4%	-11.5%	-20.6%
C	-2.7%	-3.5%	-6.3%	-11.4%	-11.6%	-21.4%
D	-2.3%	-3.7%	-6.8%	-12.1%	-10.9%	-22.1%
E	-3.1%	-2.9%	-5.9%	-11.5%	-11.3%	-20.7%
All	**-2.7%**	**-3.6%**	**-6.6%**	**-11.4%**	**-11.5%**	**-21.2%**

It is important to note that the number of iterations is an important parameter that determines the trade-off at the encoder side between the compression efficiency performance and the encoding time, which typically varies in different use cases such as live and Video-on-Demand (VoD) streaming. To put the used values in perspective, our experiments show that on average, 130%, 220% and 410% encoder complexity in terms of run-time have been imposed for 1-iter, 5-iter and 20-iter settings, respectively.

Fig. 4 presents a visual assessment of the proposed method. In this example, the two faces in the *KristenAndSara* sequence are treated as the ROI in the proposed method. Three settings are evaluated, namely 1) Base coder *i.e.* no refinement, 2) Base refinement *i.e.* no rate allocation and 3) proposed refinement with rate allocation. In all settings the bitrates of the image are kept almost the same and the only difference is in their quality. When comparing the base coder and the base refinement, it can be seen that on avareage, the PSNR of the ROI regions is increased about 0.47 dB. It is important to note that the PSNR increase of regions outside the ROI is also in the same range as the they were treated in a similar way by the base refinement. However, when comparing the proposed refinement with the base coder, on average, the PSNR increase of the ROI is around 1.1 dB, while the PSNR of the regions outside the ROI is either unchanged or slightly deteriorated.

V. Conclusion

This paper presents a method that enables learning-based image/video encoders apply spatial rate allocation with arbitrary interest map. To do so, a post-encoding latent refinement process is integrated at the frame level in the encoder in order to find an alternative latent representation of the frame which better preserves pixels of interest. As this method is agnostic to the underlying encoder as well as the algorithm that determines the regions of interest, it can be arbitrary used for different purposes. Performance evaluation on top an existing LVC shows the proposed method is able to save bitrate at the same level of the weighted PSNR, which is aligned with the pixel interest map. As next step, one can integrate the same method in an actual video coding setting. That is to say refining all types of frames (*i.e.* I, P and B) in different Group of Picture (GOP) structure. Moreover, it is also possible to conduct formal subjective quality assessment tests to validate how much the proposed rate allocation method improves the subjective quality perception.

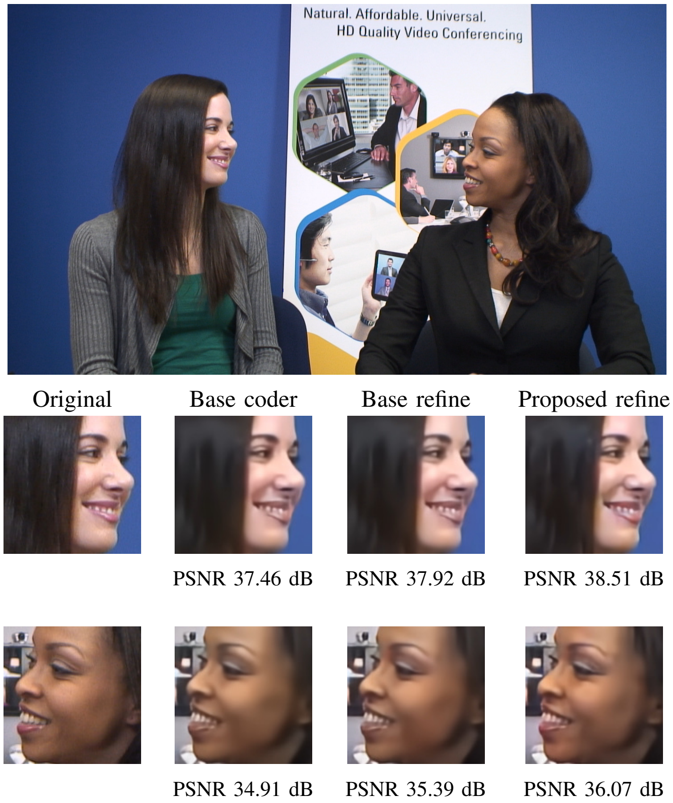

Fig. 4: Visual performance assessment on the first frame of the KristenAndSara sequence, where the two faces are used as the region-of-interest.

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/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8503–8512.

[2] Amirhossein Habibian, Ties van Rozendaal, Jakub M Tomczak, and Taco S Cohen, "Video compression with rate-distortion autoencoders," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7033–7042.

[3] Reza Pourreza and Taco Cohen, "Extending neural p-frame codecs for b-frame coding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6680–6689.

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

[5] 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.

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

[7] Chunlei Cai, Li Chen, Xiaoyun Zhang, and Zhiyong Gao, "End-to-end optimized roi image compression," *IEEE Transactions on Image Processing*, vol. 29, pp. 3442–3457, 2019.

[8] Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, and David Zhang, "Learning convolutional networks for content-weighted image compression," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3214–3223.

[9] 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*, 2021, pp. 2380–2389.

[10] 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)*. IEEE, 2020, pp. 1–6.

[11] Fabian Brand, Kristian Fischer, and André Kaup, "Rate-distortion optimized learning-based image compression using an adaptive hierachical autoencoder with conditional hyperprior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1885–1889.

[12] Fabian Brand, Kristian Fischer, Alexander Kopte, Marc Windsheimer, and André Kaup, "Rdonet: Rate-distortion optimized learned image compression with variable depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1759–1763.

[13] Fabian Brand, Kristian Fischer, Alexander Kopte, and André Kaup, "Learning true rate-distortion-optimization for end-to-end image compression," *arXiv preprint arXiv:2201.01586*, 2022.

[14] Noor Fathima, Jens Petersen, Guillaume Sautière, Auke Wiggers, and Reza Pourreza, "A neural video codec with spatial rate-distortion control," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5365–5374.

[15] Yadong Lu, Yinhao Zhu, Yang Yang, Amir Said, and Taco S Cohen, "Progressive neural image compression with nested quantization and latent ordering," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 539–543.

[16] Théo Ladune, Gordon Clare, Pierrick Philippe, and Félix Henry, "Artificial Intelligence based Video Codec (AIVC) for CLIC 2022," in *CLIC 2022, 5th Workshop and Challenge on Learned Image Compression, CVPR 2022*, 2022.

[17] Gisle Bjontegaard, "Calculation of average psnr differences between rd-curves," *ITU SG16 Doc. VCEG-M33*, 2001.

[18] Johannes Erfurt, Christian R Helmrich, Sebastian Bosse, Heiko Schwarz, Detlev Marpe, and Thomas Wiegand, "A study of the perceptually weighted peak signal-to-noise ratio (wpsnr) for image compression," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2339–2343.

[19] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, and Qingyao Wu, "A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection," *arXiv preprint arXiv:2203.04708*, 2022.