CAESR: Conditional Autoencoder and Super-Resolution for Learned Spatial Scalability

Charles Bonnineau⋆†‡, Wassim Hamidouche⋆‡, Jean-François Travers†, Naty Sidaty†,
Jean-Yves Aubié⋆ and Olivier Deforges‡

⋆IRT b<>com, Cesson-Sevigne, France,
†TDF, Cesson-Sevigne, France,
‡Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France

Abstract—In this paper, we present CAESR, an hybrid learning-based coding approach for spatial scalability based on the versatile video coding (VVC) standard. Our framework considers a low-resolution signal encoded with VVC intra-mode as a base-layer (BL), and a deep conditional autoencoder with hyperprior (AE-HP) as an enhancement-layer (EL) model. The EL encoder takes as inputs both the upscaled BL reconstruction and the original image. Our approach relies on conditional coding that learns the optimal mixture of the source and the upscaled BL image, enabling better performance than residual coding. On the decoder side, a super-resolution (SR) module is used to recover high-resolution details and invert the conditional coding process. Experimental results have shown that our solution is competitive with the VVC full-resolution intra coding while being scalable.

Index Terms—Spatial Scalability, Conditional Autoencoder, Super-Resolution, VVC

I. INTRODUCTION

Over the last years, spatial scalability has been considered as a key challenge for image and video compression. Hence, dedicated video coding standards have been developed to take advantage of the existing correlations between different versions of a signal. In the case of scalable high efficiency video coding (SHVC) [1], a base-layer (BL) signal (low resolution) encoded with high efficiency video coding (HEVC) is used as a reference by an inter-layer processing module to encode the enhancement-layer (EL) signal (high-resolution) with the use of high level syntax (HLS). More recently, low complexity enhancement video coding (LCEVC) [2] proposed specific tools to encode the residual information, i.e., the difference between the original video and its compressed representation. In these approaches, all tools, including scaling and transform modules, are handcrafted and separately tuned. Therefore, they may result in a suboptimal system.

Another way to enable spatial scalability relies on spatial resolution adaptation coding framework. In this coding scheme, illustrated at the bottom of Fig. 1, a downscaled representation of the source signal is encoded, transmitted, and upscaled after decoding to reach the original resolution. At low-bitrate, this process may provide better coding performance than full-resolution coding [3] while enabling spatial scalability with any base-layer codec. With the recent advances in deep learning, powerful pre and post-processing models have been used for spatial resolution adaptation based on existing compression standards [4]–[7]. However, some high frequencies lost during the downscaling process still cannot be recovered using single post-processing modules, making performance sensitive to the content.
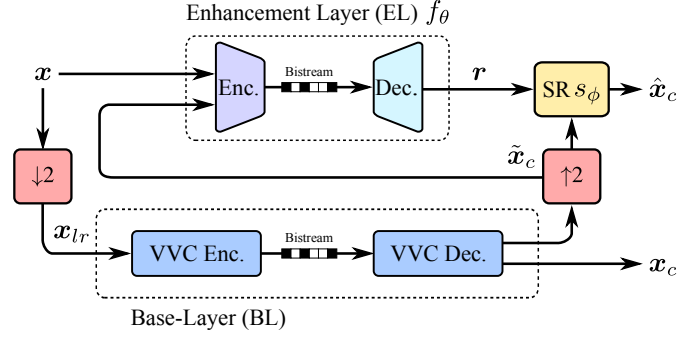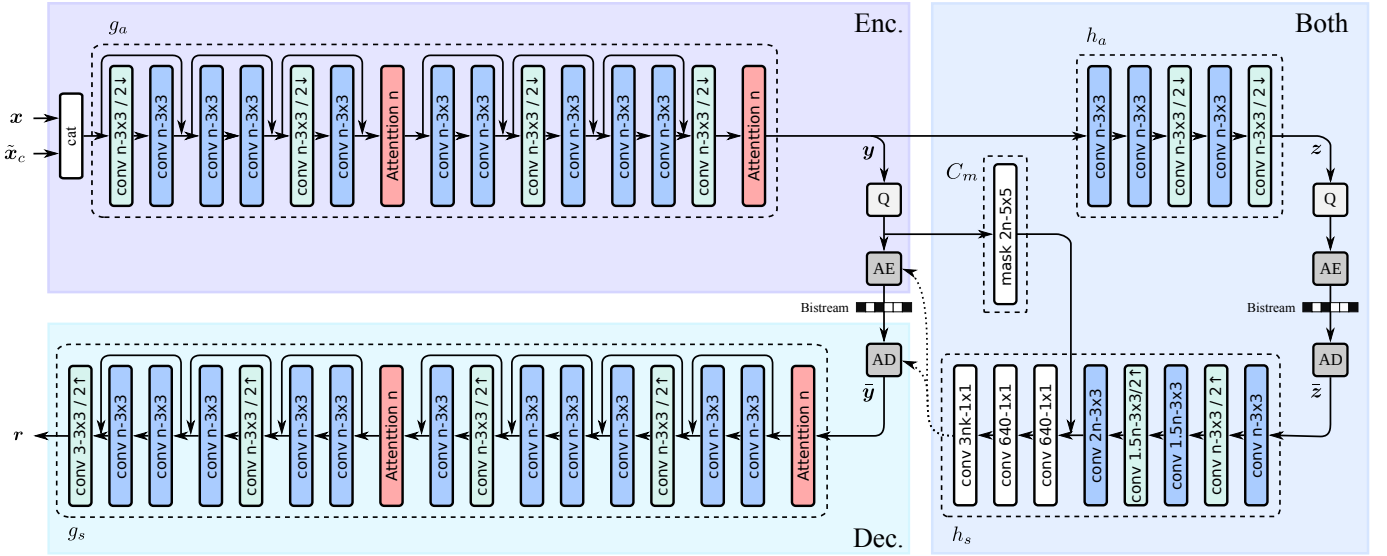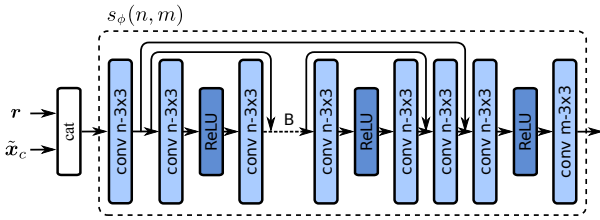

Fig. 1: General pipeline of CAESR. Both the downscaling and upscaling steps, denoted as ↓2 and ↑2, respectively, are performed by a handcrafted filter. On the decoder side, it allows matching both the latent residual information r and the upscaled base-layer signal $\tilde{x}_c$ resolutions as input of the super-resolution (SR) module s_ϕ.

On the other hand, end-to-end learning models for image and video compression were proposed using deep autoencoders (AEs) [8]–[12]. These deep models consist of a non-linear encoder-decoder pair optimized in a completely end-to-end fashion. Thus, the whole system's components are optimally tuned together regarding a given rate-distortion trade-off driven by the loss function. Hybrid layered systems have been investigated to enhance traditional codecs using an AE as an enhancement layer model [13]–[15]. However, those solutions are based on full resolution BL images and do not take into consideration the spatial scalability character.
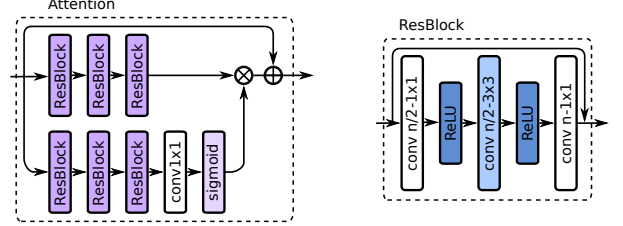
In this paper, we present CAESR, an hybrid layered approach that uses a downscaled representation of the input image, encoded using versatile video coding (VVC) as a BL codec and a deep conditional autoencoder as an enhancement-layer (EL) model. The key idea is to use the strong representation ability of AEs to encode the high-resolution details lost during the downscaling and quantization steps. On the decoder side, the predicted residual information is given with the upscaled BL signal as input to a super-resolution module to produce the reconstructed high-resolution image. We optimize the overall system by training the autoencoder jointly with the super-resolution convolutional neural network (CNN) in a conditional coding scheme. To the best of our knowledge, no previous works consider spatial scalability based on the joint training of a super-resolution module and an autoencoder to transmit both coding and scaling residuals as side information.

(a) Architecture details of the conditional autoencoder f_θ. g_a and g_s correspond to the main encoder and decoder, h_a and h_s to the hyper-encoder and hyper-decoder, and C_m to the autoregressive context model described in [16]. Skip connections represent element-wise additions between features. Q, AE and AD stand for quantization, arithmetic encoding, and arithmetic decoding steps, respectively. We fix $n = 192$.

(b) Architecture details of the super-resolution network s_ϕ. Skip connections stand for element-wise additions between features.

(c) Building blocks of the conditional autoencoder f_θ. These attention and residual blocks are implemented as proposed in [11].

Fig. 2: Architecture and details of CAESR.

II. PROPOSED SOLUTION

A. Framework and Formulation

The overall pipeline of the proposed solution is described in Fig. 1. This hybrid layered system takes a low-resolution signal encoded with VVC as a base-layer and a conditional autoencoder with hyperprior (AE-HP) [10] as an enhancement-layer that feeds a learning-based SR module. In the following, let s_ϕ denotes the SR module and f_θ the parametric function of the conditional autoencoder with hyperprior.

Given an input image $\boldsymbol{x} \in \mathbb{R}^{W \times H \times 3}$ of width W and height H, we first apply a spatial downscale by a factor 2 to generate the BL images $\boldsymbol{x}_{lr}$. This latter is encoded with a VVC encoder. The decoded image $\boldsymbol{x}_c$ is then rescaled to the original resolution $W \times H$ to form the EL model's input $\tilde{\boldsymbol{x}}_c$.

Our approach relies on conditional coding that allows a non-linear mixture of the source and the reconstructed BL signal to be learned, improving the performance compared to residual coding [17]. Thus, the source image $\boldsymbol{x}$ and the upscaled base reconstruction $\tilde{\boldsymbol{x}}_c$ are concatenated along the feature axis to feed the autoencoder. The resulting tensor $(\tilde{\boldsymbol{x}}_c, \boldsymbol{x}) \in \mathbb{R}^{W \times H \times 6}$ is encoded by the encoder part of f_θ, denoted as g_a, into a latent vector $\boldsymbol{y}$. Additional latent variables $\boldsymbol{z}$ are produced by the hyper-encoder h_a to capture spatial dependencies among

the element of $\boldsymbol{y}$. Both latents are quantized using the $round$ function to produce $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{z}}$. At training, we apply a uniform noise $\mathcal{U}(-\frac{1}{2}, +\frac{1}{2})$ on latents to emulate the quantization errors while enabling backpropagation, resulting in $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{z}}$. To simplify, we use $\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{z}}$ to denote both actual and emulated quantized latents. The latent variables are then entropy coded regarding a gaussian mixture model (GMM) parameterized by the output of the hyper-decoder h_s as:

$$p(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}}) \sim \sum_{k=1}^{K} \boldsymbol{w}^{(k)} \mathcal{N}(\boldsymbol{\mu}^{(k)}, \boldsymbol{\sigma}^{2(k)}), \qquad (1)$$

with k the index of mixtures defined by $\boldsymbol{w}^{(k)}$, $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\sigma}^{2(k)}$, denoting weights, means and scales, respectively.

At the decoder side, the latent residual signal r is reconstructed by the synthesis part of f_θ, denoted as g_s, and concatenated with the upscaled based-layer image $\tilde{\boldsymbol{x}}_c$ to form the input of the super-resolution network s_ϕ. Finally, the output image $\hat{\boldsymbol{x}}_c$ is reconstructed from the following equation:

$$\hat{\boldsymbol{x}}_c = s_\phi(\tilde{\boldsymbol{x}}_c, r). \qquad (2)$$

In this work, the upscaling operation is applied before feeding the super-resolution module s_ϕ using an interpolation

filter to make the network performing both high-resolution details recovering and conditional coding process inversion.

All components of the overall differentiable system are jointly trained to minimize the following rate distortion loss function $\mathcal{L}$ based on a Lagrangian multiplier λ:

$$\mathcal{L}(\lambda) = D(\hat{\boldsymbol{x}}_c, \boldsymbol{x}) + \lambda R. \tag{3}$$

The distortion D is measured using the mean squared error (MSE) between $\hat{\boldsymbol{x}}_c$ and $\boldsymbol{x}$. The term R corresponds to the Shannon entropy of $\tilde{\boldsymbol{y}}$, computed as:

$$R = \mathbb{E}_{\tilde{\boldsymbol{y}} \sim m}[-\log_2(p(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{z}}))], \tag{4}$$

with m the true distribution of latents.

B. Network Architecture

The architecture of the proposed system, illustrated in Fig. 2, is described in this section.

1) Autoencoder: The structure of f_θ is based on the layered autoencoder with hyperprior (AE-HP) architecture described in [11], that estimates the group of parameters $\{\boldsymbol{w}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\sigma}^{2(k)}\}$, with $k = 3$, for the entropy model described in (1). We also use an autoregressive context model over latents [16], denoted as C_m, to improve the entropy model accuracy without increasing the rate. The main analysis and synthesis transforms, g_a and g_s, respectively, are composed of successive self-attention and residual blocks, depicted in Fig. 2c. The non-linearity is integrated using the generalized divisive normalization (GDN) activation function [9] and LeakyReLU as described in [11]. For the hyper-encoder h_a and hyper-decoder h_s, LeakyReLU activation function is used. Regarding dimensionality reduction and expansion strided convolutional layers and sub-pixel upscaling layers [18] are implemented, respectively.

2) Super-Resolution: Our super-resolution module is inspired by the enhanced deep super-resolution (EDSR) architecture [19] which enables state-of-the-art performance. This SR architecture mainly consists of B residual blocks (RBs) with short and long skip connections. In this work, we fix $B = 8$ and use 64 filters of size 3×3 for each convolutional layer. We introduce the non-linearity with the ReLU activation, as described in Fig. 2b. We removed the upscaling layer typically located at the end of the network and perform image upscaling before passing the input picture through this module.

III. Experimental Results

A. Training

Both super-resolution and autoencoder networks are jointly trained to minimize the rate-distortion loss defined in (3).

We train our model using three different image datasets, namely DIV2K [20], Flickr2K, and the training dataset provided by the challenge on learned image compression (CLIC21) [21]. The performance is evaluated on the CLIC21 validation dataset, consisting of 42 images with various spatial characteristics. We first convert the image samples from PNG to YUV4:2:0 format. The base-layer input images $\boldsymbol{x}_{lr}$ are obtained by applying a bicubic downscale of factor 2. Then, the reconstructed versions $\tilde{\boldsymbol{x}}_c$ of the low-resolution images $\boldsymbol{x}_{lr}$, are obtained using the VVC test model (VTM-11) in all-intra configuration for different quantization parameters (QPs). For simplicity, we generate YUV4:4:4 tensors by duplicating the chroma components for both reconstructed images $\tilde{\boldsymbol{x}}_c$ and original images $\boldsymbol{x}$, respectively. We crop 256×256 high-resolution and corresponding 128×128 low-resolution patches from the training set, resulting in around 150K training pairs.

We train one model per base-layer $QP \in \{37, 32, 27, 22\}$ and select specific λ values in (3). As the base quality is starting to saturate at higher bitrate, we empirically decided to allocate more bitrate for the lower BL QPs. The models are trained over a total of 20 epochs with a learning rate of 10^{-4}. We apply a learning rate decay with a gamma of 0.5 for the last 5 epochs to improve the convergence. We use a batch size of 8 and optimize the model with ADAM [22] by setting $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. For the whole experiments, the quality is assessed on the luma component using peak signal to noise ratio (PSNR) and structural similarity (SSIM) [23] full-reference objective image quality metrics computed between the reconstructed images $\hat{\boldsymbol{x}}_c$ and original images $\boldsymbol{x}$.

B. Ablation Study

In this experiment, we demonstrate the effectiveness of the proposed system through an ablation study. The models that use our EL module f_θ, including the proposed conditional coding system CAESR and the residual-based configurations with and without super-resolution, represented by res_{sr} and res_{bic}, respectively, are illustrated in Fig. 3. For those configurations, we compute the global rate by summing both the BL and EL signal bitrates. This test also considers configurations based on our super-resolution module s_ϕ and a bicubic interpolation filter used as post-processing modules, represented by sr and bic, respectively. The whole learned models are optimized using the training strategy described in Section III-A.

We display latent variables and bitmaps obtained with the different tested models in the right part of Fig. 3. We observe that the configurations that include super-resolution, i.e., (a) and (b), produce more sparse latent variables that require fewer bits for enhancement layer encoding. The joint training of the super-resolution module s_ϕ and the autoencoder f_θ allows an optimal interaction between the two models. Therefore, high-frequencies that can be recovered by super-resolution are omitted by the autoencoder, allowing the autoencoder f_θ to focus on the most complex areas.

The rate-distortion (RD) curves are represented in Fig. 4. We also add full-resolution single layer VVC configuration, which corresponds to the high-resolution images encoded with VVC VTM-11 all-intra mode. To match the bitrates obtained with our layered system, we select $QP \in \{42, 39, 36, 31\}$ for full-resolution coding. The proposed conditional system outperforms all the other tested configurations in terms of rate-distortion performance, using the BD-BR (Bjøntegaard-Delta Bit-Rate) metric [24], on the whole bitrate range. While offering spatial scalability, the BD-BR values of CAESR
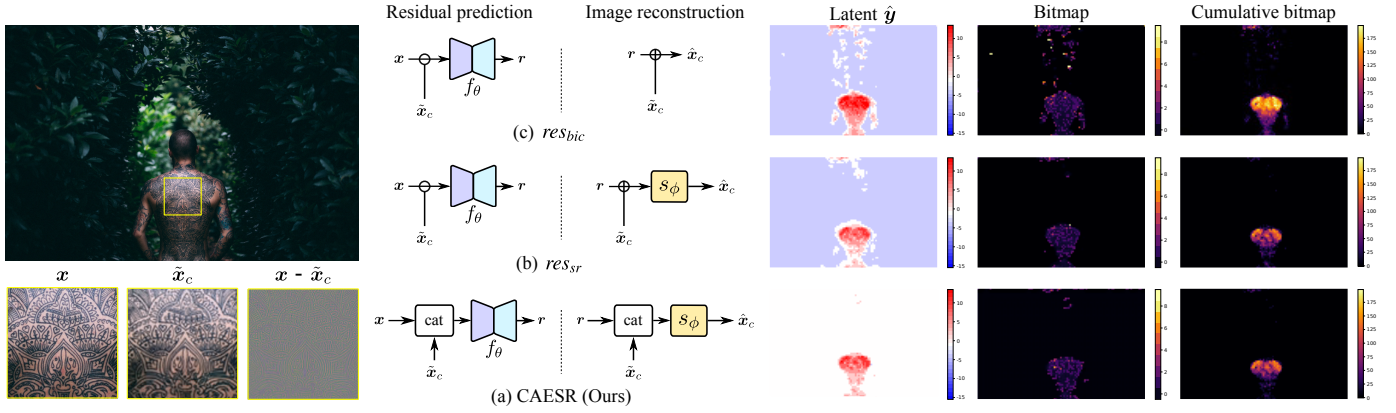
Fig. 3: Visualization of the configurations tested during ablation that include the autoencoder f_θ using *felix-russell-saw-140699.png* from the CLIC21 validation set encoded with VTM-11 AI (qp27) as an example. We illustrate residual prediction and image reconstruction steps for each configuration. The latent $\hat{y}$ and its corresponding bitmap represent the channel with the highest entropy. Cumulative bitmap corresponds to the sum of bitmaps over the latent channels.
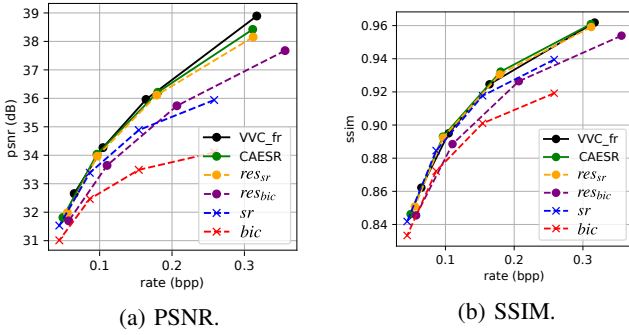

(a) PSNR.   (b) SSIM.

Fig. 4: Performance of the tested configurations on the CLIC2021 validation image dataset. We display incomplete systems for ablation study in dashed lines.

over the full-resolution coding anchor are 3.41% and -3.49% regarding the PSNR and SSIM metrics, respectively. We notice that the configurations that include both the autoencoder f_θ and the SR module s_ϕ in the enhancement layer are more efficient than the others, particularly at higher bitrates. Indeed, in this range of bitrate, the reconstructed residual information contains high-resolution details that cannot be recovered using a single post-processing module. Although the residual bicubic configuration, i.e., (c) in Fig. 3, offers lower performance, this experiment demonstrates that at high bitrate, simply transmitting the residual with our system offers gains in PSNR over super-resolution used as a post-processing module.

C. Visualization

In this experiment, we visually compare our method against the super-resolution network EDSR [19] used as a post-processing module on images. To allow a fair evaluation, we trained EDSR following the experimental settings described in Section III-A. For visualisation, we adjust the QP of the EDSR input to match the bitrate with our system.

As depicted in Fig. 5, our method produces better high-resolution images in terms of visual quality than both the bicubic filter and EDSR used as post-processing, while being close to the full-resolution coding anchor. We observe that

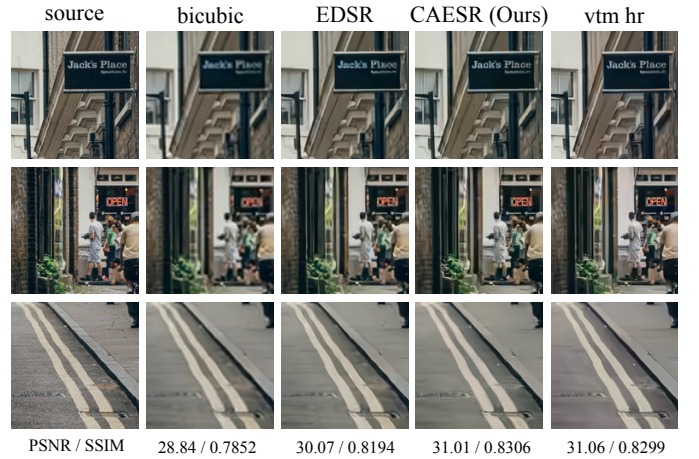|	source	bicubic	EDSR	CAESR (Ours)	vtm hr
PSNR / SSIM		28.84 / 0.7852	30.07 / 0.8194	31.01 / 0.8306	31.06 / 0.8299

Fig. 5: Visual illustration of *daniel-robert-405.png* (0.21bpp).

our system allows highly contrasted areas, like texts, to be accurately recovered from the low-resolution image.

IV. CONCLUSION

In this paper, we present CAESR, an hybrid learning-based approach for spatial scalability based on the joint training of two deep convolutional neural networks (CNNs): a conditional autoencoder f_θ and a super-resolution module s_ϕ. The deep autoencoder with hyperprior, learns to represent the residual information that cannot be recovered by the super-resolution module used as a post-processing step. This residual information is combined with the upscaled base-layer reconstruction at the decoder side to form the high-resolution output signal. Our approach relies on conditional coding that learns the optimal mixture of the source and the upscaled image, enabling better performance than residual coding. Our solution offers performances on par with VVC full-resolution intra coding while being scalable.

As future work, we plan to include the temporal aspect into our model to ensure inter-coded frame processing.

REFERENCES

[1] J. M. Boyce, Y. Ye, J. Chen, and A. K. Ramasubramonian, "Overview of shvc: Scalable extensions of the high efficiency video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 20–34, 2015.

[2] F. Maurer, S. Battista, L. Ciccarelli, G. Meardi, and S. Ferrara, "Overview of mpeg-5 part 2–low complexity enhancement video coding (lcevc)," *ITU Journal: ICT Discoveries*, vol. 3, no. 1, 2020.

[3] A. M. Bruckstein, M. Elad, and R. Kimmel, "Down-scaling for better transform compression," *IEEE Transactions on Image Processing*, vol. 12, no. 9, pp. 1132–1144, 2003.

[4] F. Zhang, M. Afonso, and D. R. Bull, "Vistra2: Video coding using spatial resolution and effective bit depth adaptation," *Signal Processing: Image Communication*, p. 116355, 2021.

[5] D. Ma, M. Afonso, F. Zhang, and D. R. Bull, "Perceptually-inspired super-resolution of compressed videos," in *Applications of Digital Image Processing XLII*, vol. 11137. International Society for Optics and Photonics, 2019, p. 1113717.

[6] C. Bonnineau, W. Hamidouche, J.-F. Travers, and O. Deforges, "Versatile video coding and super-resolution for efficient delivery of 8k video with 4k backward-compatibility," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2048–2052.

[7] C. Bonnineau, W. Hamidouche, J.-F. Travers, N. Sidaty, and O. Deforges, "Multitask learning for vvc quality enhancement and super-resolution," in *2021 Picture Coding Symposium (PCS)*, 2021, pp. 1–5.

[8] G. Toderici, S. M. O'Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar, "Variable rate image compression with recurrent neural networks," in *International Conference on Learning Representations*, 2016. [Online]. Available: http://arxiv.org/abs/1511.06085

[9] "End-to-end optimized image compression," 2017, 5th International Conference on Learning Representations, ICLR 2017 ; Conference date: 24-04-2017 Through 26-04-2017.

[10] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rkcQFMZRb

[11] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7939–7948.

[12] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "Dvc: An end-to-end deep video compression framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 006–11 015.

[13] Y.-H. Tsai, M.-Y. Liu, D. Sun, M.-H. Yang, and J. Kautz, "Learning binary residual representations for domain-specific video streaming," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[14] M. Akbari, J. Liang, and J. Han, "Dsslic: deep semantic segmentation-based layered image compression," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2042–2046.

[15] W.-C. Lee, C.-P. Chang, W.-H. Peng, and H.-M. Hang, "A hybrid layered image compressor with deep-learning technique," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2020, pp. 1–6.

[16] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in Neural Information Processing Systems*, vol. 31, pp. 10 771–10 780, 2018.

[17] T. Ladune, P. Philippe, W. Hamidouche, L. Zhang, and O. Déforges, "Conditional coding for flexible learned video compression," in *International Conference on Learning Representations (ICLR) 2021, Neural Compression Workshop*, 2021.

[18] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[19] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[20] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 126–135.

[21] C. on Learned Image Compression, "https://www.compression.cc/," June 2021.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[24] G. Bjøntegaard, "Document VCEG-M33 ITU-T Q6/16: Calculation of Average PSNR Differences Between RD- Curves," April 2001.