An Effective Entropy Model for Semantic Feature Compression

Tianma Shen, Ying Liu†
Department of Computer Science and Engineering, Santa Clara University, USA
Emails: tshen2@scu.edu, yliu15@scu.edu

***Abstract*—Semantic feature compression aims to compress image features for downstream machine vision tasks without reconstructing image pixels. Such a task is very challenging since it needs to learn features which are not only useful for machine vision tasks, but also easy to compress. While existing learnable feature coding models utilize downstream task networks as teacher networks to guide the learning and compression of semantic features, they use simple entropy models and do not effectively reduce information redundancy. In this work, we propose a transformer-based spatial-channel auto-regressive feature context model (SC-AR FCM) to assist the entropy coding of learnable features. Through extensive experimentation on object detection and segmentation tasks, we demonstrate that the rate-accuracy performance of our proposed method surpasses traditional image compression techniques and state-of-the-art learning-based feature compression techniques.**

***Index Terms*—context model, entropy model, image coding for machines, object detection, segmentation**

I. Introduction

In recent years, the rapid advancement of machine vision applications has placed an unprecedented demand on the compression of images. These applications, ranging from image classification [1] and object detection [2] to semantic segmentation [3], rely on the analysis at a scale never before imagined. To meet this demand, a new technological paradigm, image coding for machines (ICM) [4]–[12] has emerged. ICM helps in efficiently handling various vision tasks by compressing the information contained in images, allowing machines to process visual data more effectively.

Existing ICM methods can be categorized into two approaches: compress-then-analyze methods [8], [9], [11], [12], and analyze-then-compress methods [4]–[7], [10]. Compress-then-analyze methods concatenate the image codec and the machine task network in an end-to-end manner. They need to first reconstruct image pixels. Analyze-then-compress methods extract semantic features from images and then compress those features without reconstructing image pixels [10]. Compress-then-analyze methods face a problem because machines don't need all feature maps from the original images for their vision tasks to work [8], [11], leading to a bigger bit rate. On the contrary, analyze-then-compress methods focus on compressing necessary feature maps [10] as semantic features, which have essential and relevant information for semantic vision tasks.

However, despite the promise of analyze-then-compress methods, we have identified a limitation in state-of-the-art (SOTA) methods. Most notably, the entropy models used in these methods have weaknesses that make semantic features more redundant [4]–[7], [10]. The entropy model is crucial because it predicts the probability distribution of encoded semantic features, helping to reduce bit rates in the coding process. However, current models use a simple hyperprior architecture [13] that struggles to capture correlations between semantic feature elements to estimate the probability distribution [10].

Consequently, in this paper, we focus on analyze-then-compress methods, composing their semantic feature for machines. We introduce a novel approach to address limitations of the entropy model. Our contributions are summarized as follows:

- We designed a transformer-based spatial-channel auto-regressive feature context model (SC-AR FCM) to enhance the entropy coding of semantic features.
- We evaluate our model on two machine vision tasks: objective detection and semantic segmentation. Experimental results demonstrate that our proposed method showcase state-of-the-art rate-accuracy performance.

II. Related Work

A. Compress-then-Analyze Methods

Several existing compress-then-analyze methods [8], [9], [11], [12] have explored the development of end-to-end models tailored for machine-vision tasks. An early approach, end-to-end learnable network [8], directly concatenated the image codec with a machine task network. However, the challenge we faced was finding the right balance among various loss functions, such as those for machine vision tasks, image distortion, and bit rate. To address this, end-to-end optimized network [11] introduced a pre-trained codec to reduce image distortion losses and maintain a balance in losses. The semantics-oriented network [12] took a step further by enhancing the loss function. It introduced a new perspective called semantics-oriented metrics, which effectively highlighted the importance-weighted pixels for specific machine tasks, resulting in improved performance. In addition, content-adaptive methods [9]

†Corresponding author.
This work is supported in part by the National Science Foundation under Grant ECCS-2138635 and the NVIDIA Academic Hardware Grant.

focused on refining the latent representation of end-to-end learned image codecs. This was achieved through fine-tuning the encoder to optimize the overall performance of the system.

B. Analyze-then-Compress Methods

In contrast to compress-then-analyze methods, analyze-then-compress methods [4]–[7], [10] have emerged as promising alternatives, compressing semantic features, which contain crucial and relevant information for semantic vision tasks. An early approach, deep feature compression [10] introduced an idea of extracting semantic features through machine task's encoder and then compressing these semantic features. However, this direct compression poses a challenge to learning semantic information. Compressible features codec [7] offered a new perspective by focusing on a learnable encoder to extract a latent representation instead of semantic features. Then, a learnable decoder decompresses semantic features for supervised machine tasks. Entropic student [4] adopted knowledge distillation principles to enhance the learnable encoder and decoder, by employing both a teacher model and a student model with the guidance of a teacher model to train student's encoder in advance. Efficient entropic student [5] not only improved the training process but also introduced an efficient encoder architecture with residual blocks to improve entropic student [4]'s performance. Prompt-ICM [6] directly compresses semantic features with the mask as additional input from an information selector (IS). IS generates additional information, which is relevant to supervised machine tasks, to help the codec to compress semantic features.

These analyze-then-compress methods [4]–[7], [10] makes great contributions on main decoder and decoder. However, they overlook the importance of incorporating entropy models to perform the critical task of encoding semantic features. They just utilize the simple hyperprior to estimate latent representation's distributions, which struggle to capture the full correlation of semantic feature elements, resulting in suboptimal compression efficiency for machine tasks. In the subsequent sections of this paper, we will delve deeper into these limitations and present our approach, the transformer-based spatial-channel auto-regressive feature context model (SC-AR FCM), as a solution to overcome these challenges.

III. PROPOSED METHOD

A. The Overall Architecture

Fig. 1 presents the overall architecture of our proposed learnable semantic feature compression framework for machine vision tasks. Following entropic student [4], this architecture consists of a teacher network (Fig. 1 upper pipeline) and a student network (Fig. 1 lower pipeline), which includes our proposed entropy model. The teacher network is a machine task network, such as an object detection net or a semantic segmentation net, which does not have data compression capability. It guides and instructs the training of the student network. The backbone of the teacher network is ResNet50, which has five blue stage blocks shown in Fig. 1. The task head at the end of the teacher network is to address

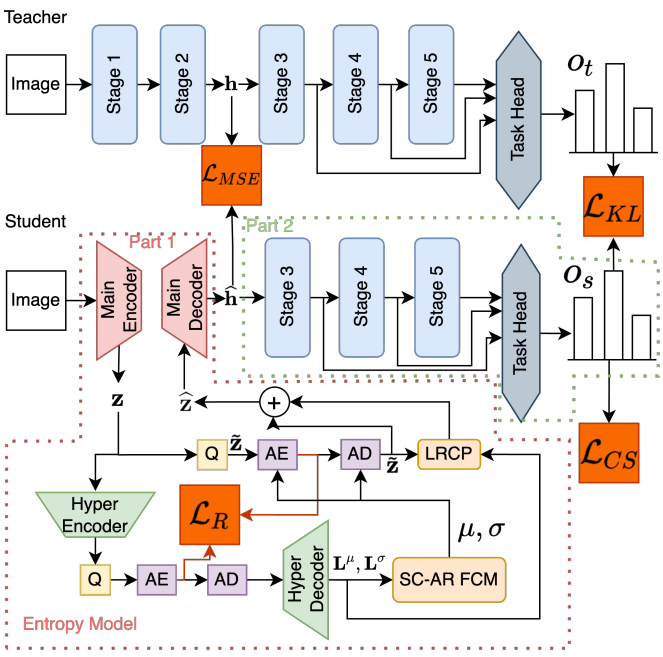

Fig. 1. The overall architecture of the proposed model. AE and AD are the arithmetic encoder and arithmetic decoder. Q represents quantization. SC-AR TCM represents a spatial-channel auto-regressive feature context model. LRCP represents a latent residual cross-attention prediction.

specific machine vision tasks. We have selected RetinaNet [2] for object detection and Deeplab V3 [3] for semantic segmentation, employing these as our teacher networks.

The student network, in contrast, is divided into two fundamental components, Part 1 and Part 2 as shown in Fig. 1. Part 1 is the feature codec, which plays a pivotal role in learning and compressing semantic features useful for the downstream machine task. The decoded semantic feature $\hat{\mathbf{h}}$ is expected to match the Stage 2 output feature $\mathbf{h}$ of the teacher network. During the training process, the teacher network provides essential guidance to the student network in learning semantic features, by minimizing the mean squared error (MSE) between $\hat{\mathbf{h}}$ and $\mathbf{h}$, and minimizing the Kullback-Leibler (KL) divergence between $\mathbf{o}_t$ and $\mathbf{o}_s$, which are the soft labels (class probabilities) output by the teacher network and the student network, respectively. In order to reduce the redundancy of semantic features, we proposed spatial-channel auto-regressive feature context model (SC-AR FCM). SC-AR FCM captures both spatial and channel correlations within the semantic latent representation by estimating μ and σ for the latent representation's distribution in entropy coding. Subsequently, Part 2 of the student network replicates the remaining pipeline of the teacher network, and finally outputs the machine task results, such as the object detection bounding boxes and the semantic segmentation maps.

B. The Main Encoder/Decoder and Hyper Encoder/Decoder

Our proposed student network has a feature codec as Part 1 shown in Fig. 1 that effectively extracts and compresses semantic features. The feature codec consists of the main

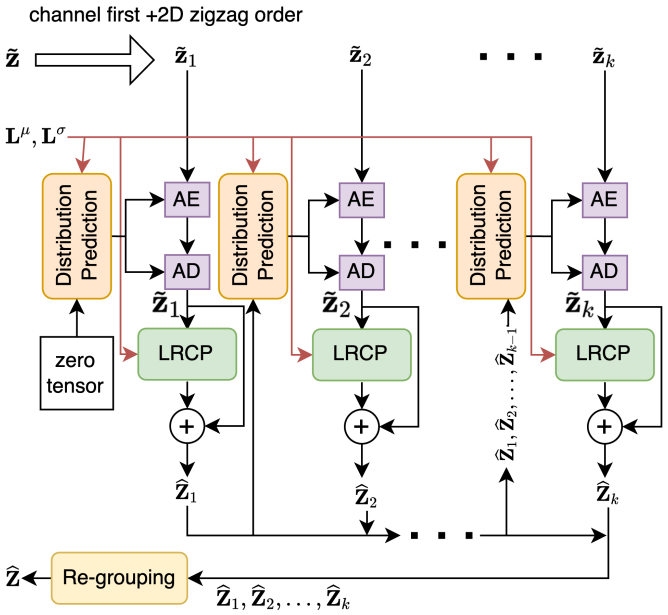

Fig. 2. The architecture of spatial-channel auto-regressive context model. LRCP represents a latent residual cross-attention prediction.

encoder/decoder, the hyper encoder/decoder, our SC-AR FCM, and latent residual cross-attention prediction (LRCP). The main encoder/decoder consists of four convolutional layers for encoding and decoding, while the hyper encoder/decoder employs three convolutional layers. The main encoder directly extracts semantic features from the input image $\mathbf{x}$ and compresses them as a latent representation $\mathbf{z}$, which is quantized as $\widetilde{\mathbf{z}}$, and further processed using an arithmetic encoder (AE) to create a bitstream. The bitstream is then sent to an arithmetic decoder (AD) to recover $\widetilde{\mathbf{z}}$. We predict a residue to reduce the quantization error of $\widetilde{\mathbf{z}}$ by LRCP (details in Section C). Once we compute this residue, we add it to $\widetilde{\mathbf{z}}$ to obtain $\widehat{\mathbf{z}}$ which is finally decoded by the main decoder to recover the semantic features $\widehat{\mathbf{h}}$.

In the entropy model, the hyper encoder and decoder assist the proposed SC-AR FCM in estimating the distributions of $\widetilde{\mathbf{z}}$ for AE and AD. These components work together to ensure efficient entropy coding, ultimately contributing to the success of our proposed architecture.

C. SC-AR FCM and LRCP

Fig. 2 shows the details of our proposed spatial-channel auto-regressive feature context model (SC-AR FCM) for entropy coding. In our pursuit of reducing redundancy within $\widetilde{\mathbf{z}}$ for machines, we employ a SC-AR FCM. This model focuses on learning both spatial and channel correlations within the semantic latent representation $\widetilde{\mathbf{z}}$, such that entropy coding will be more efficient and require less bit rates. As depicted in Fig. 2, $\widetilde{\mathbf{z}}$ is divided into small chunks, denoted as $\widetilde{\mathbf{z}}_1, \widetilde{\mathbf{z}}_2, \ldots, \widetilde{\mathbf{z}}_k$ along both the spatial and channel dimensions. These chunks are entropy coded in a channel-first, then spatial 2D zigzag order. Our proposed distribution prediction module estimates the distribution parameters μ_k and σ_k for $\widetilde{\mathbf{z}}_k$, using the outputs of the hyper decoder $\mathbf{L}^{\mu}$ and $\mathbf{L}^{\sigma}$ and all previously decoded chunks $\widetilde{\mathbf{z}}_1, \widetilde{\mathbf{z}}_2, \ldots, \widetilde{\mathbf{z}}_{k-1}$.

When dealing with semantic features, it is crucial to mitigate quantization errors to maintain the quality of the decoded feature maps. To address this concern, we introduce a LRCP designed to estimate the decimal part of $\mathbf{z}$. Both distribution prediction module and LRCP are constructed using a shifted-window transformer [14], enabling them to effectively capture global dependencies within feature maps. Once all the chunks, namely $\widetilde{\mathbf{z}}_1, \widetilde{\mathbf{z}}_2, \ldots, \widetilde{\mathbf{z}}_k$, have been decoded, they are reassembled into a coherent whole, forming the reconstructed $\widehat{\mathbf{z}}$. This process ensures that the essential information for machine tasks is efficiently preserved while minimizing redundancy and quantization errors.

D. Training and Loss Functions

For the semantic segmentation task, we train our proposed model in an end-to-end manner with the following loss function $\mathcal{L}_{\text{seg}}$,

$$\mathcal{L}_{\text{seg}} = \lambda \cdot \mathcal{L}_R + 0.1 \cdot \mathcal{L}_{MSE} + 0.5 \cdot \mathcal{L}_{KL} + \mathcal{L}_{PCE}, \quad (1)$$

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N_h} \|\mathbf{h} - \widehat{\mathbf{h}}\|_2^2, \quad (2)$$

$$\mathcal{L}_{KL} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{M} o_t^{c,n} \log \left(\frac{o_s^{c,n}}{o_t^{c,n}} \right), \quad (3)$$

$$\mathcal{L}_{PCE} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{M} y_c \log \left(o_t^{c,n} \right). \quad (4)$$

Here, $\mathcal{L}_R$ represents the bit rate of $\widetilde{\mathbf{z}}$ which is measured by bits per pixel (BPP), $\mathcal{L}_{MSE}$ accounts for the mean squared error between the semantic feature of the teacher network $\mathbf{h}$ and that of the student network $\widehat{\mathbf{h}}$, N_h is the number of elements in $\mathbf{h}$, N is the number of pixels in an image, $\mathcal{L}_{KL}$ is the KL divergence between $\mathbf{o}_s$ and $\mathbf{o}_t$, $\mathcal{L}_{PCE}$ corresponds to the pixel-level cross-entropy loss, c is the class index, M is number of classes, and $y_c \in \{0, 1\}$ is the label of class c, using 1-of-M coding.

For the object detection task, we choose (5) as the loss function.

$$\mathcal{L}_{\text{obj}} = \lambda \cdot \mathcal{L}_R + 0.1 \cdot \mathcal{L}_{MSE} + 0.5 \cdot \mathcal{L}_{KL} + \mathcal{L}_{BCE} + \frac{1}{N_b} \mathcal{L}_b \quad (5)$$

$$\mathcal{L}_{BCE} = -\frac{1}{N_b} \sum_{n=1}^{N_b} \sum_{c=1}^{M} y_c \log \left(o_t^{c,n} \right), \quad (6)$$

$$\mathcal{L}_b = \sum_{i=0}^{N_{\text{box}}-1} \|\mathbf{p}_i - \widehat{\mathbf{p}}_i\|_2. \quad (7)$$

Here, $\mathcal{L}_{BCE}$ corresponds to the bounding box's cross-entropy loss, $\mathcal{L}_b$ corresponds to intersection-over-unions (IoU) loss of bounding boxes, N_b is the number of bounding boxes, whose IoU is higher than 0.5, $\mathbf{p}$ represents the ground-truth bounding box coordinates, and $\widehat{\mathbf{p}}$ represents the bounding box coordinates predicted by the student network.

By adopting this unified training approach, we aim to overcome the limitations stemming from the frozen main encoder, providing a more effective and streamlined solution for our model's training process.

IV. EXPERIMENTS AND ANALYSIS

A. Datasets and Evaluation Metrics

In our experiments, we employed the COCO 2017 dataset [15] for both object detection and semantic segmentation tasks. This dataset is well-regarded in the computer vision community for its comprehensive collection of images and annotations.

Throughout our training process, we continuously assessed our model's performance on the validation dataset. For each training epoch, we closely monitored the model's loss on this validation dataset. If the model's loss on the validation dataset ceased to decrease, we terminated the training process.

When evaluating the performance of our model in the object detection task, we employed the metric of mean average precision (mAP). This metric is calculated based on bounding box (BBox) outputs and considers different Intersection-over-Unions (IoU) thresholds, ranging from 0.5 to 0.95.

For the semantic segmentation task, our performance assessment relied on mIoU value, which is averaged IoU across 21 distinct segment classes. This approach provided a detailed understanding of our model's segmentation accuracy, offering insights into its ability to differentiate and segment objects within the images.

B. Object Detection Results

To assess the effectiveness of our proposed model, we conducted an experiment against the state-of-the-art method, Entropic Student [4], and also conventional codecs, VTM-19.2 [16] and BPG [17]. Notably, both VTM-19.2 and BPG belong to the category of compress-then-analyze methods, wherein they decompress source images for machine input.

In the context of object detection, our results, as illustrated in Fig. 3 (a), showcase the performance of our model in terms of rate-distortion curves and mean average precision (mAP). Notably, our model exhibits superior rate-distortion curves, reflecting its ability to efficiently balance the compression rates and the preservation of object detection accuracy. Of particular significance is the performance at low bit rates, where our model achieves an mAP that surpasses VTM-19.2, the second-best method, by an impressive margin of 8.9%. This improvement at lower bit rates highlights the robustness and efficiency of our approach.

C. Semantic Segmentation Results

As shown in Fig. 3 (b), we conducted a comparative analysis of our model's performance with respect to previous methods for the semantic segmentation task. Notably, our model exhibits higher performance than other methods, especially at lower bit rates. Our model surpasses the entropic student approach, achieving a 0.21% increase in mIoU. This outcome underscores the effectiveness of our proposed method

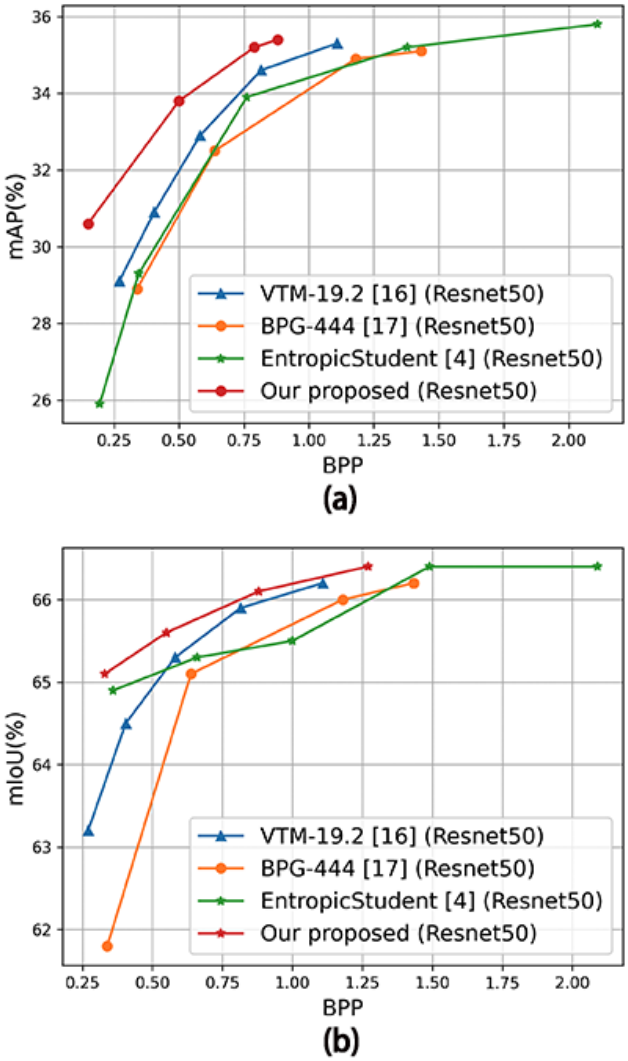

Fig. 3. The figure (a) is the results of object detection task. The figure (b) is the results of semantic segmentation task.

in accurately semantic features from images. Moreover, in a comparison with VTM, our model also demonstrates an improvement in mIoU, boasting a 1.17% increase. This performance enhancement shows the capacity of our approach to excel in semantic segmentation.

V. CONCLUSION

In this paper, we propose the transformer-based spatial-channel auto-regressive feature context model (SC-AR FCM), which enhances the entropy coding of learnable features for machine vision tasks. Our extensive experiments in object detection and segmentation showcase that our method outperforms conventional image compression techniques and the state-of-the-art feature compression model, achieving superior rate-accuracy performance. In the future, we will extend our

framework to accommodate multiple machine vision tasks or hybrid machine-human vision tasks.

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Miami, Florida, USA, June 20-25, 2009, pp. 248–255.

[2] T. Lin, P. Goyal, R. B. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proc. IEEE International Conference on Computer Vision, Venice, Italy, October 22-29, 2017, pp. 2999–3007.

[3] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587 (2017).

[4] Y. Matsubara, R. Yang, M. Levorato, S. Mandt, Supervised compression for resource-constrained edge computing systems, in: Proc. IEEE/CVF Winter Conference on Applications of Computer Vision, IEEE, Waikoloa, HI, USA, January 3-8, 2022, pp. 923–933.

[5] Z. Duan, F. Zhu, Efficient feature compression for edge-cloud systems, in: Picture Coding Symposium (PCS), IEEE, San Jose, CA, USA, Dec 2022, pp. 187–191.

[6] R. Feng, J. Liu, X. Jin, X. Pan, H. Sun, Z. Chen, Prompt-icm: A unified framework towards image coding for machines with task-driven prompts, arXiv preprint arXiv:2305.02578 (2023).

[7] S. Singh, S. Abu-El-Haija, N. Johnston, J. Ballé, A. Shrivastava, G. Toderici, End-to-end learning of compressible features, in: Proc. IEEE International Conference on Image Processing, IEEE, Abu Dhabi, United Arab Emirates, October 25-28, 2020, pp. 3349–3353.

[8] N. Le, H. Zhang, F. Cricri, R. G. Youvalari, E. Rahtu, Image coding for machines: an end-to-end learned approach, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, June 6-11, 2021, pp. 1590–1594.

[9] N. Le, H. Zhang, F. Cricri, R. G. Youvalari, H. R. Tavakoli, E. Rahtu, Learned image coding for machines: A content-adaptive approach, in: Proc. IEEE International Conference on Multimedia and Expo, Shenzhen, China, July 5-9, 2021, pp. 1–6.

[10] Z. Chen, K. Fan, S. Wang, L. Duan, W. Lin, A. C. Kot, Toward intelligent sensing: Intermediate deep feature compression, IEEE Transactions on Image Processing 29 (2020) 2230–2243.

[11] L. D. Chamain, F. Racapé, J. Bégaint, A. Pushparaja, S. Feltman, End-to-end optimized image compression for machines, a study, in: Proc. Data Compression Conference, Snowbird, UT, USA, March 23-26, 2021, pp. 163–172.

[12] C. Gao, D. Liu, L. Li, F. Wu, Towards task-generic image compression: A study of semantics-oriented metrics, IEEE Transactions on Multimedia 25 (2023) 721–735.

[13] D. Minnen, J. Ballé, G. Toderici, Joint autoregressive and hierarchical priors for learned image compression, in: S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Proc. Annual Conference on Neural Information Processing Systems, Montréal, Canada, December 3-8, 2018, pp. 10794–10803.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proc. IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, October 10-17, 2021, pp. 9992–10002.

[15] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: common objects in context, in: Proc. Computer Vision European Conference, Vol. 8693, Zurich, Switzerland, September 6-12, 2014, pp. 740–755.

[16] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, J.-R. Ohm, Overview of the versatile video coding (vvc) standard and its applications, IEEE Trans. on Circuits and Systems for Video Technology 31 (10) (2021) 3736–3764.

[17] F. Bellard, BPG image format, `http://bellard.org/bpg/` (Accessed: 2022-1-18).