Rethinking Semantic Image Compression: Scalable Representation With Cross-Modality Transfer

Pingping Zhang, Shiqi Wang, *Senior Member, IEEE*, Meng Wang, Jiguo Li, Xu Wang, *Member, IEEE*, and Sam Kwong, *Fellow, IEEE*

***Abstract*— This article proposes the *scalable cross-modality compression* (SCMC) paradigm, in which the image compression problem is further cast into a representation task by hierarchically sketching the image with different modalities. Herein, we adopt the conceptual organization philosophy to model the overwhelmingly complicated visual patterns, based upon the semantic, structure, and signal level representation accounting for different tasks. The SCMC paradigm that incorporates the representation at different granularities supports diverse application scenarios, such as high-level semantic communication and low-level image reconstruction. The decoder, which enables the recovery of the visual information, benefits from the scalable coding based upon the semantic, structure, and signal layers. Qualitative and quantitative results demonstrate that the SCMC can convey accurate semantic and perceptual information of images, especially at low bitrates, and promising rate-distortion performance has been achieved compared to state-of-the-art methods. The code will be available online https://github.com/ppingzhang/SCMC.**

***Index Terms*— Semantic image compression, cross-modality, scalable coding.**

I. Introduction

RECENT years have witnessed the exciting development of machine learning technologies, which make the fully data driven image compression solutions become possible [1], [2], [3]. The main objective of image compression is to maximize the ultimate utility of the reconstructed visual information, given the constrained number of used bits. Central to such a problem is the way in which the images can be finally utilized. With the advance of computer vision, the images, which excel at conveying the visual information, can be understood and perceived in a variety of ways. The semantic information, which is intrinsically critical in image understanding, plays an important role in the visual information representation. In particular, it enjoys several advantages, including being compact to represent, friendly to understand, as well as closely tied to visual signals. However, the semantic information has been unfortunately ignored in current learning based image representation models, in particular when the end-to-end coding strategy converts the visual signals to the latent code without sufficient interpretability. Li et al. [4] proposed a cross-modal compression framework to achieve semantic communication, but this approach essentially preserves semantic consistency while the signal-level reconstruction has not been fully considered.

Scalable compression has been proven to be an efficient representation method by encoding the visual signals into several layers [5], [6], [7], [8], [9]. As such, the decoding of higher layers typically relies on the existence of lower layers [7], [8], [9], [10]. More specifically, the compact feature representation and visual signal compression can be naturally incorporated into a unified scalable coding framework, based upon the excellent reconstruction capability of deep learning models. Herein, we propose the scalable cross-modality

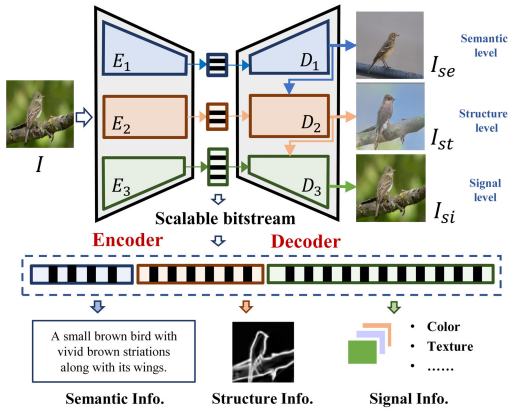

Fig. 1. The paradigm of SCMC. The hierarchical organization based upon the semantic, structure and signal level representation forms the SCMC. Specifically, the encoder and decoder consist of three layers, including the semantic layer (E_1 and D_1), structure layer (E_2 and D_2) and the signal layer (E_3 and D_3). These layers work coherently and seamlessly in SCMC.

Manuscript received 4 August 2022; revised 25 October 2022 and 22 December 2022; accepted 18 January 2023. Date of publication 31 January 2023; date of current version 4 August 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62022002 and Grant 61871270, in part by the Shenzhen Science and Technology Program under Project JCYJ20220530140816037, in part by the Shenzhen Natural Science Foundation under Grant JCYJ20200109110410133, in part by the Hong Kong Innovation and Technology Commission (InnoHK) through Project Centre for Intelligent Multidimensional Data Analysis (CIMAD), and in part by the Hong Kong General Research Fund-Research Grants Council (GRF-RGC) under Grant 11209819 (CityU 9042816) and Grant 11203820 (9042598). This article was recommended by Associate Editor T. Lu. *(Corresponding author: Shiqi Wang.)*

Pingping Zhang and Meng Wang are with the Department of Computer Science, City University of Hong Kong, Hong Kong, China (e-mail: ppingyes@gmail.com; mwang98-c@my.cityu.edu.hk).

Shiqi Wang and Sam Kwong are with the Department of Computer Science, City University of Hong Kong, Hong Kong, China, and also with the City University of Hong Kong, Shenzhen Research Institute, Shenzhen 518057, China (e-mail: shiqwang@cityu.edu.hk; cssamk@cityu.edu.hk).

Jiguo Li is with the Institute of Computing Technology, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the School of Information Science and Technology, Fudan University, Shanghai 200437, China (e-mail: jiguo.li@vipl.ict.ac.cn).

Xu Wang is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China, and also with the Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen University, Shenzhen 518060, China (e-mail: wangxu@szu.edu.cn).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCSVT.2023.3241225.

Digital Object Identifier 10.1109/TCSVT.2023.3241225

compression scheme (SCMC), which transfers the visual signals into different modalities for representation. As shown in Fig. 1, the SCMC scheme plays a bridging role between semantic image understanding and image representation, with the three specifically designed layers. First, extremely compact representation can be achieved with the base layer that encodes semantic information only. Second, between semantic understanding and visual signal reconstruction, geometric structures (e.g., edges and ridges) are extracted, bridging their gap to satisfy diverse demands. Such representation is essentially based upon Marr's Theory on the computational representation framework [11], and lays the foundation for providing a perceptually meaningful representation. Third, the signal-level reconstruction is enabled in the third layer, enhancing the robustness of the proposed compression scheme and providing the faithful reconstruction with sufficient signal-level details.

The proposed SCMC enjoys several desired advantages, including compact, flexible and high efficiency. In particular, when only semantic information is required, the proposed framework enables a natural base-layer representation with very compact information conveyed. Moreover, each layer of the scalable stream holds conceptually meaningful information, enhancing the flexibility by decoding the subset layers for a given task and enhancing the interpretability of the bitstream. Finally, the interactions between different layers and fusions of different layers, which are indispensable in scalable representation, ensure the promising rate-distortion performance.

The contributions of this paper can be summarized as follows,

1) We propose a novel SCMC framework based upon the semantic, structure and signal layers. Qualitative and quantitative results demonstrate that our proposed SCMC can convey accurate semantic, structure and signal level visual information with diverse configurations, significantly promoting the compression performance.
2) We design the three layers representation in SCMC, following the conceptual organization and coherent design philosophy in the scalable coding framework. The three layers sequentially recover the visual information at semantic, structure and signal levels, and such representation architecture is expected to make profound impacts on a broad range from image processing to image understanding.
3) We develop the interaction strategy among the three layers, ensuring that the decoding of higher layers is supported by the existence of lower layers. As such, the redundancy among layers can be efficiently removed in the scalable image representation paradigm based upon the solutions in aligning and fusing cross-modality features.

II. The Proposed SCMC Framework

The proposed SCMC framework contains an encoder and a decoder. Both of them consist of three layers, including the semantic layer, structure layer and signal layer, to support high-level semantic communication and low-level image reconstruction. The detailed design is shown in Fig. 1. The encoder separately extracts semantic, structure, and signal level representation via layered compression, yielding embedded bitstreams. More specifically, the base layer is the semantic layer, compressing the image data into the captions to convey the semantic information with an ultra-low bitrate. The structure layer, as the second layer, extracts the structure information which is further compressed with the VVC [12]. The signal layer serves as the final layer, compressing signal representation based upon the existing learning-based codec [1].

In the decoder, these bitstreams can be partially decoded to obtain visual reconstruction from semantic, structure and signal perspectives. The semantic layer reconstructs the semantic information from compact text descriptions. The structure layer generates images by decompressing semantic bitstream and structure bitstream, promoting the perceptual reconstruction of images. The signal layer as the final signal level reconstruction is intrinsically based upon the reconstructed images from the first two layers. The information from the previous layer serves as conditional information, such that this interaction strategy ensures that redundancy among layers can be efficiently removed, leading to scalable cross-modality image compression.

A. Semantic Layer

The semantic information extraction based upon the image captioning lays the foundation for the base layer, which could be represented in an extremely compact way. As such, instead of extracting the semantic information at the receiver given the corrupted image from the decoder, the proposed scheme allows the high quality reconstruction of the semantic information even at ultra-low bitrate. Moreover, based upon the text-to-image (T2I) generation, the visual signals with the same semantic information can be generated from the base layer, although the signal level reconstruction cannot be guaranteed. Thus, the heart of the base layer lies in image-text-image (ITI) cross-modality translation, compression and representation. More specifically, it is composed of three submodules, including the image-to-text (I2T) translation, lossless compression for text description, and the T2I generation. In this paradigm, the I2T translation in the encoder aims to compress the data from the signal domain into a compact text description. Herein, we utilize an end-to-end neural network which is capable of automatically generating a reasonable description in plain English [13]. In comparison to the visual signals, the text domain is semantically meaningful and compact. However, statistical redundancy still exists such that Huffman coding [14] is employed to remove statistical redundancy in text compression. The T2I generation in the decoder aims to reconstruct visual images with semantic consistency, such that we leverage AttnGAN [15] to reconstruct images from the text descriptions, providing the semantically similar visual information.

B. Structure Layer

Following the insight of Marr's theory [11], geometric structures (e.g., edges and ridges) and stochastic textures are two prominent components composing the visual scene. As such, we compress the structure map of the input image I into a bitstream with low bitrates and reconstruct the image I_{st}

based on structures and semantic textures. In the encoder, the structure information is extracted, and then compressed into the bitstream.

In the decoder, we leverage a combined reconstruction scheme of geometric structures and semantic textures from the base layer to improve representation capability. This layer contains three stages, including structure extraction and compression, structure-semantic layer fusion, and image reconstruction.

1) Structure Extraction and Compression: In the encoder, the structural map of the input image I is obtained through the Richer Convolutional Features (RCF) structure extraction [16]. RCF can fully exploit multiscale and multilevel information of objects to encapsulate both semantic and fine detail features, such that structure extraction could be both accurate and efficient. More importantly, even though the structure map extracted via RCF is compressed with a high compression ratio, it is still able to maintain a good structure. To compactly represent the structural information, we first downsample structure maps by a factor of 2 and compress the downsampled structural map via VTM with QP 50 under all intra (AI) configurations, wherein the screen content coding (SCC) tools [17] are enabled, due to the strong capability in compressing screen content images with sharp and abundant edges. Finally, we can obtain the bitstream of the structure maps. On the decoder side, the reconstructed structure map I_e and semantic texture map I_{se} are combined to facilitate the generation of the perceptual reconstruction of this layer.

2) Structure-Semantic Layer Fusion: This operation contains two stages, including aligning structure and semantic features and fusing the aligned structure and semantic features. Due to information inconsistency between semantically generated texture from the base layer and structure, we convert texture and structure maps into feature domains to align them via a multi-scale alignment strategy.

After aligning the structure and semantic features, structure features are merged into aligned features after self-calibrated convolution [18] via the element-wise addition. Then, an upsampling module, including a convolution operation and a PixelShuffle operation [19] are performed. Following feature upsampling, the spatial dimension is enlarged two times. For better reconstruction of the details, the semantic information and structure features are fused after adjusting them via self-calibrated convolution for further improvements. After self-calibrate convolution, the structure features perform the pixel-wise addition with aligned features to obtain fusion features as the input of the next upsampling operation. Through hierarchical calibration and fusion, we can obtain more accurate semantic texture and structure features to contribute to image generation.

3) Image Reconstruction: After a multi-scale fusion operation, the final reconstruction consists of two upsampling operations and two residual blocks, where the residual block is only composed of two convolution layers. As such, we can generate an image with similar semantics and structure as the input image.

To obtain perceptual reconstruction even at low bit rates, we design a loss function to train the structure layer. The generator G generates the image on the condition of the semantic maps I_{se} and structure maps I_e. The discriminator is then trained to distinguish the generated image $I_{st} = G(I_{se}, I_e)$ with the original image I. We train the network with LSGANs [20] in an end-to-end manner.

To maintain the semantic consistency and optimize visual quality, we introduce a new term, the DISTS [21] loss ($\mathcal{L}_{DISTS}$), to further enhance the connection between the input image (I) and the reconstructed image (I_{st}). With the enforcement of the $\mathcal{L}_1$ and $\mathcal{L}_{DISTS}$, the intrinsic similarity between the input images and the generated images is largely improved, facilitating the conceptual representations.

$$\mathcal{L}_{re} = \lambda_1 \mathcal{L}_1(I, I_{st}) + \lambda_d \mathcal{L}_{DISTS}(I, I_{st}). \tag{1}$$

As such, the objective function of the proposed framework is

$$G^* = \arg\min_G \max_D \mathcal{L}_d(G(I_{se}, I_e), D(I)) + \lambda_g \mathcal{L}_g(G(I_{se}, I_e)) + \mathcal{L}_{re}, \tag{2}$$

where $\mathcal{L}_d$ and $\mathcal{L}_g$ are the discriminative and generative loss [20]. λ_g, λ_1 and λ_d are the weighting parameters to balance each component, and we empirically set $\lambda_g = 1$, $\lambda_1 = 10$ and $\lambda_d = 10$.

After the end-to-end training, the structure layer combined with the structure features extracts the texture information from semantic images to promote image generation.

C. Signal Layer

Involving signal-level attributes (e.g., color and background) is conducive to reconstructing original image signals. In the signal layer, we concentrate on representing the signal-level information. More specifically, the signal-level information is delicately extracted from the input image I and compressed as the bitstream at the encoder-side, conveying signal-level characteristics. The decoder parses the bitstream, generating the reconstructed image I_{si} with the assistance of the associated structure information from the second layer. The framework is constructed based on an existing learning-based codec [1]. In particular, it contains an encoding module E_3, quantization (Q), entropy coding (AE/AD) and decoding module D_3. The encoder module and entropy coding module share the identical backbone with the existing learning-based codec [1], showing the promising capability of image compression.

To obtain genuine signal representation, we propose to improve the decoder by involving the initial structure-level information in the image reconstruction during the decoding process. The multi-scale structure features serve as the conditional information in the decoder. More specifically, multi-scale structure features are extracted from the decoded structure maps I_e and the output of the structure layer I_{st} via the Sobel operator. These structure features provide the layout and detailed texture information to facilitate image reconstruction. Subsequently, these structure features readjust via self-calibrated convolution and fuse with signal features through the fusion operation, which is identical to the fusion block in the structure layer. In this manner, the conditional information from the previous layer can be fully utilized to promote signal compression performance.

TABLE I
QUANTITATIVE RESULTS AT THE ULTRA-LOW BITRATE

	bpp	FID (↓)	IS(↑)
VTM-15.2	0.013	328.715	1.150
Ours-Semantic layer	**0.005**	**52.258**	**2.260**

The rate-distortion (RD) loss function ($\mathcal{L}_{RD}$) in this layer includes the content reconstruction distortion $\mathcal{L}_{mse}$ and the bitrate (R) for the image encoding [1], which is given by,

$$\mathcal{L}_{RD} = \lambda \mathcal{L}_{mse} + R, \tag{3}$$

where λ is the hyper-parameter to control the trade-off between the bitrate and distortion.

III. EXPERIMENTS

A. Datasets

We adopt CUB-200-2011 dataset [22], which includes 200 bird species. The CUB-200-2011 is adopted due to the fact that it is a popular dataset with caption information, which could greatly facilitate the proposed task. The dataset is divided into training and testing subsets. The training dataset includes 8855 images with 160 bird species and the testing dataset contains 2933 images with 40 bird species. Each image is associated with 10 descriptions. In the following experiments, the images are resized to 256×256.

B. Quality Evaluation Measures

Peak signal and noise rate (PSNR) is a widely-used metric in compression and restoration tasks, We employ LPIPS [23] and DISTS [21] as the quality evaluation measures. In particular, a lower DISTS/LPIPS value indicates better quality. The coding bitrate is evaluated as the bits per pixel (bpp).

C. Experimental Settings

The network is implemented in the PyTorch framework and trained on NVIDIA GeForce RTX 3090 GPUs. We provide detailed information regarding the experimental settings of three layers.

1) Semantic Layer: This stage contains two training steps, including training the I2T translation and the T2I generation. For the I2T translation, we set the batch size to 128 and the learning rate to 0.001 with 100 epochs. Images are randomly cropped to 224×224. Other settings follow those in [13]. For the T2I generation, we follow the settings of AttnGAN [15].

2) Structure Layer: We set the batch size to 16 and the learning rate to 0.0001 with 200 epochs. Moreover, regarding the compression of the structure maps, we adopt the VVC test model (VTM-15.2) [12] of screen content coding (SCC) under AI configuration, where the QP is set as 50.

3) Signal Layer: We employ the learning-based codec, Ballé et al. [1], as the backbone. We set the batch size to 128 and the learning rate to 0.001 with 200 epochs. The λ is set as 5×2^t, where t is equal to $\{0, 2, 4, 6, 8\}$, corresponding to different bitrate points.

D. Performance Comparisons

To verify the effectiveness of the proposed SCMC scheme, the following image compression schemes are involved for performance comparisons,

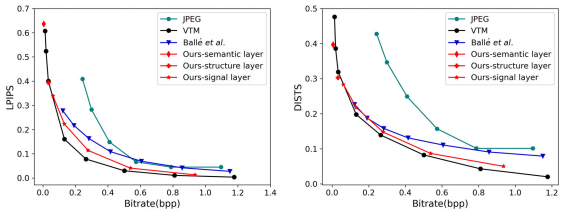

Fig. 2. Comparisons of the R-D performance wherein the LPIPS and DISTS are employed as quality evaluation measure.

- **JPEG**: we use JPEG encoder with the quality factors QFs = {1, 5, 10, 20, 30, 40}, corresponding to the compression ratios from large to small.
- **VVC (Intra)**: we employ the VVC test model (VTM-15.2) [12] with quantization parameters QPs = {63, 57, 52, 42, 37, 32, 27, 22}, and higher QP corresponds to lower bitrate.
- **Ballé et al.'s method** [1]: the training and testing strategies follow those provided by CompressAI [24].

To evaluate the compression performance of the proposed framework quantitatively, we compare the proposed ITI with the JPEG, VTM, and Ballé et al.'s method. We compress all the images in the testing set with different quality factors. The Rate-Distortion (RD) performance comparisons are illustrated in Fig. 2.

1) Compression Performance of the Semantic Layer: The proposed semantic layer can achieve ultra-high compression ratios with semantically promising texture reconstructions. However, extremely low-bitrate compression is difficult to attain when employing JPEG and Ballé et al.'s framework. As shown in Fig. 2, "Ours-semantic layer" denotes the R-D performance of the semantic layer in terms of the LPIPS and DISTS. It is easy to observe that the proposed method has close performance with VTM at the ultra-low bitrate, whereas the proposed method outperforms VTM when evaluated with DISTS. Meanwhile, we evaluate the quality of generated semantic images on FID and IS, as shown in Table I. The proposed model with lower bitrate compression can still bring better FID and IS performance.

2) Compression Performance of the Structure Layer: The structure layer compresses image data with the assistance of the semantic texture and structure maps. The comparison results of the structure layer are shown in Fig. 2, where "Ours-structure layer" illustrates the R-D performance of the structure layer in terms of the LPIPS and DISTS. The results indicate the advantage of our method at low bitrates.

3) Compression Performance of the Signal Layer: The signal layer is responsible for conveying signal-level visual information with enhanced reconstructions. The quantitative results regarding the R-D performance and visualization results are shown in Fig. 2 and Fig. 3, respectively. The proposed signal layer surpasses JPEG and Ballé et al.'s inferences of LPIPS and DISTS. Fig. 3 illustrates the decoded images via JPEG, Ballé et al.'s method, VTM, and Ours from left to right. We can clearly observe blocking artifacts and color shifts when using JPEG compression. Moreover, the reconstructed images are blurred when employing Ballé et al.'s method. Regarding VVC, images compressed with VTM exhibit noticeable blocking artifacts in the background regions. By contrast, owing to the cooperation of the structure information, the proposed

Fig. 3. Visual quality comparison on the decoded images. The values below each image are bpp and DISTS values, where the lower DISTS value represents the better reconstruction quality.

model provides satisfied visual quality with similar or even smaller coding bits. Moreover, in terms of PSNR, the proposed method is also superior to Ballé et al.'s method with 20.6% BD-rate gains, though inferior to VVC (50.7% loss). This is not surprising that the VVC focuses on the signal level recovery during encoder optimization. More results can be found in the supplementary material.

IV. Conclusion

In this paper, we have presented a novel SCMC framework where a wide spectrum of novel functionalities has been enabled, making the codec versatile for applications ranging from semantic understanding to signal-level reconstruction. The proposed layered bitstream can be truncated due to the scalability design, and ideally, such a rate-scalable method could meet the demands of diverse requirements. The proposed coding architecture is intrinsically hierarchical, and promising coding performance has been shown in a variety of means, demonstrating the promise of SCMC in real-world applications.

Herein, we took a certain viewpoint regarding how the images could be represented by a variety of ways, ranging from semantic level representation to signal level reconstruction. The message we are trying to send is not that the proposed paradigm is superior to existing methods all through. Rather, we hope to make the point that the proposed paradigm could be an alternative but effective solution for specific application domains. Moreover, though useful evidences have been provided on the effectiveness of the cross-modality coding paradigm, there remain spaces for further exploration. In particular, it is highly expected that the coding paradigm can adapt to various image contents with different resolutions, and even be enhanced from the perspectives of scalability, interoperability, utility, and feasibility to meet the grand challenges of compact data representation in a variety of visual-centered applications.

References

[1] J. Ballé, D. Minnen, S. Singh, S. Jin Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," 2018, *arXiv:1802.01436*.

[2] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, "Image and video compression with neural networks: A review," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1683–1698, Jun. 2020.

[3] Y. Wang, D. Liu, S. Ma, F. Wu, and W. Gao, "Ensemble learning-based rate-distortion optimization for end-to-end image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1193–1207, Mar. 2021.

[4] J. Li, C. Jia, X. Zhang, S. Ma, and W. Gao, "Cross modal compression: Towards human-comprehensible semantic compression," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 4230–4238.

[5] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[6] A. Heindel, E. Wige, and A. Kaup, "Low-complexity enhancement layer compression for scalable lossless video coding based on HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1749–1760, Aug. 2017.

[7] K. Liu, D. Liu, L. Li, N. Yan, and H. Li, "Semantics-to-signal scalable image compression with learned revertible representations," *Int. J. Comput. Vis.*, vol. 129, no. 9, pp. 2605–2621, Sep. 2021.

[8] S. Wang et al., "Towards analysis-friendly face representation with scalable feature and texture compression," *IEEE Trans. Multimedia*, vol. 24, pp. 3169–3181, 2022.

[9] C. Cai, L. Chen, X. Zhang, G. Lu, and Z. Gao, "A novel deep progressive image compression framework," in *Proc. Picture Coding Symp. (PCS)*, Nov. 2019, pp. 1–5.

[10] H. Tu, L. Li, W. Zhou, and H. Li, "Semantic scalable image compression with cross-layer priors," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4044–4052.

[11] D. Marr, *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. Cambridge, MA, USA: MIT Press, 2010.

[12] *VVC Software VTM-15.2*. Accessed: Mar. 5, 2022. [Online]. Available: https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tree/VTM-15.2

[13] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, Apr. 2017.

[14] D. Huffman, "A method for the construction of minimum-redundancy codes," *Proc. IRE*, vol. 40, no. 9, pp. 1098–1101, Sep. 1952.

[15] T. Xu et al., "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.

[16] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3000–3009.

[17] W. Zhu, W. Ding, J. Xu, Y. Shi, and B. Yin, "Screen content coding based on HEVC framework," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1316–1326, Aug. 2014.

[18] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng, "Improving convolutional networks with self-calibrated convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10096–10105.

[19] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[20] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

[21] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, Dec. 2020.

[22] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011. [Online]. Available: https://www.vision.caltech.edu/datasets/cub_200_2011/

[23] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[24] J. Bégaint, F. Racapé, S. Feltman, and A. Pushparaja, "CompressAI: A PyTorch library and evaluation platform for end-to-end compression research," 2020, *arXiv:2011.03029*.