Guest Editorial
Advances in Generative Visual Signal Coding and Processing

THIS special issue of IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS (JETCAS) is dedicated to demonstrating the latest developments in algorithms, implementations, and applications related to visual signal coding and processing with generative models. In recent years, generative models have emerged as one of the most significant and rapidly developing areas of research in artificial intelligence. They have proved to be an important instrument for advancing research in AI-based visual signal coding and processing. For instance, the variational autoencoder (VAE) has been used as a fundamental framework for end-to-end learned image coding, the autoregressive (AR) model has been extensively studied for efficient entropy coding, and the generative adversarial network (GAN) has been utilized frequently to enhance the subjective quality of coding schemes. Meanwhile, generative models have also been explored in various visual signal processing tasks, including quality assessment, restoration, enhancement, editing, and interpolation.

In light of the rapid growth of generative visual signal coding and processing, its contributions to international standards and practical applications are increasingly valued. The Joint Video Experts Team (JVET) of ITU-T Video Coding Experts Group (VCEG) and ISO/IEC Moving Picture Experts Group (MPEG) started as early as 2018 an exploration study of the neural network-based video coding (NNVC) technology beyond the capabilities of the conventional hybrid video coding framework. In addition to that, MPEG and JPEG launched many standardization projects, which have started adopting AI-based technologies, such as AI-based 3-D graphics coding, AI model compression, video coding for machines (VCM), and JPEG AI. In addition, it is challenging to build a hardware accelerator that is both general and efficient to accommodate the fast-paced evolution of generative neural networks. The purpose of this Special Issue is to showcase the latest advancements in emerging technology for visual signal coding and processing with generative models.

I. ORGANIZATION AND OVERVIEW

A total of 52 submissions underwent a thorough review process and 15 papers were finally selected for publication. These accepted papers can be classified into five distinct themes: 1) generative image coding for perceptual and machine vision applications; 2) video and point cloud coding with generative models; 3) generative visual data restoration and enhancement; 4) visual content generation and analysis; and 5) FPGA implementations for generative models and learned image codecs.

We begin with a tutorial-style paper [A1]. This is a literature review authored by the Guest Editors of this Special Issue. This paper overviews the recent advancements in the field of visual signal coding and processing with generative models. It provides a concise introduction to several advanced generative models, namely the VAE, GAN, AR, and diffusion models. It further explores the advancements in visual signal coding techniques that adopt generative models, along with the relevant international standardization activities. The paper also discusses their applications and advancements in visual signal processing, including restoration, synthesis, and editing. One notable part is the utilization of generative models for visual signal quality assessment and the research on the evaluation of generative models. Last but not least, this paper presents a detailed account of fast, low-complexity, and low-power system implementations for generative visual signal coding and processing.

II. GENERATIVE IMAGE CODING FOR PERCEPTUAL AND MACHINE VISION APPLICATIONS

This group of papers focuses on generative image coding for perceptual and machine vision applications, including three papers. The first paper by Duan et al. [A2]. This paper provides an AI-generated image dataset PKU-AIGI-500K, which is composed of 105k+ text prompts and 528k+ images. To solve the domain shift between AI-generated images and natural images, a new cross-attention transformer-based neural network is proposed. Besides, text information is included in the coding framework to further improve the compression performance. Compared with H.266/VVC, up to 30.09% BD-rate saving can be achieved. In addition, the proposal also outperforms the recent learned image compression methods.

The second paper by Kong et al. [A3]. It proposes a two-layer approach for image coding. The base layer is based on a learned or traditional coding framework. To compensate for the high-frequency loss in the base layer, the enhancement layer refines the decoded image by using vector quantization. The vectorized codebook is learned and optimized for a better decoder-side refinement. The effectiveness of the method is

Digital Object Identifier 10.1109/JETCAS.2024.3403318

evaluated on both the traditional and learned compression frameworks.

The third paper by Li and Zhang [A4]. This paper raises a new question on image compression for both human perception and machine vision. The task needs to balance between the image visual quality and machine vision accuracy at a given compression ratio. This paper proposes a coding framework based on implicit neural representations (INRs) by designing a semantic embedding enhancement module to assist in understanding image semantics, thereby to enhance the model's perception of images for machine vision.

III. Video and Point Cloud Coding With Generative Models

This group of papers focuses on video and point cloud coding with generative models, including two papers.

The first paper by Du et al. [A5]. This paper proposes a contextual generative video compression method utilizing generative adversarial networks (GANs) and contextual coding to enhance perceptual quality and improve coding efficiency. By employing a hybrid transformer-convolution structure and novel entropy models, CGVC-T surpasses the state-of-the-art learned and industrial video codecs, achieving BD-rate savings in perceptual quality metrics across various datasets.

The second paper by Sun et al. [A6]. In this paper, the context feature residuals of adjacent context are considered in the new context models together with a multi-layer perception branch employed for the node occupancy prediction instead of using the probability distribution in existing works. The approach has been reported to achieve consistent bitrate savings (up to 3%) in geometry point cloud encoding at the cost of moderate computational complexity overhead.

IV. Generative Visual Data Restoration and Enhancement

This group of papers focuses on generative visual data restoration and enhancement, including four papers.

The first paper by Yan et al. [A7]. This paper addresses the problem of video inpainting by proposing a flow-guided global-local aggregation transformer network. It first employs a pre-trained optical flow complementing network to enhance the failed motion flow. A content inpainting module based on spatial and temporal transformers is used to improve the inpainting of local corrupted areas. A structural rectification module is also developed to maintain content coherence through combining both local and global features around the inpainted regions. This method has been compared with existing inpainting approaches and has demonstrated its effective performance on stabilized video content.

The second paper by Khalifeh et al. [A8]. This paper addresses video frame interpolation and proposes a multi-encoder method, which can significantly reduce model parameters while maintaining the interpolation performance. This approach specifically focuses on kernel-based video interpolation methods, and achieves consistent improvement when integrated into various network backbones.

The third paper by Zhu et al. [A9]. Based on global contextual understanding and detailed feature extraction, the proposed model can effectively merge features from various modalities, including depth and color images. This approach has been reported to offer enhanced super-resolution performance over existing approaches when evaluated on multiple RGB-D datasets.

The fourth paper by Ma et al. [A10]. The paper proposes a transformer-based network for image restoration. Different from the self-attention in Swin Transformer, an adaptive dynamic window scheme and a channel-spatial mixed attention module are proposed. Besides, to capture local dependencies, CNN is integrated in the MLP layer of Swin Transformer. The results illustrate the effectiveness on reducing VVC compression artifacts for four different QPs on three benchmark datasets.

V. Visual Content Generation and Analysis

There are three papers addressing the research topics related to visual content generation and analysis.

The first paper by Zeng et al. [A11]. It presents the first generative model (Physically Guided Generative Adversarial Network, PGGAN) to transform the multi-view light field to holographic 3D content. More specifically, this work introduces a novel framework employing an encoder-generator-discriminator architecture informed by a physical optics model, facilitating rapid learning and transfer to new scenes while adhering to holographic standards. By leveraging a differentiable physical model and adaptive loss strategy, the PGGAN achieves exceptional speed, significantly surpassing current techniques in both speed and angular reconstruction fidelity, thereby advancing real-time holographic rendering capabilities.

The second paper by Ji and Karam [A12]. This work, which focuses on learning-based compressed-domain image classification, is based on a compressed-domain vision transformer with a novel feature patch embedding capturing within- and cross-channel information in the compressed domain. Trained by an adaptation training strategy, the proposed method shows improved classification performance over existing compressed-domain classification models, with much lower computational complexity compared to pixel-domain classifiers.

The third paper by Wan et al. [A13]. In this paper, a spatio-temporal consistency generative network is proposed for 360° video saliency prediction based on a dual-stream encoder-decoder architecture with an axial-attention memory spherical ConvLSTM module. This approach has been tested on PVS-HM and VR-Eyetracking databases and demonstrates enhanced video saliency prediction performance over existing methods.

VI. FPGA Implementations for Generative Models and Learned Image Codecs

This group of papers focuses on FPGA implementations for generative models and learned image codec, including two papers.

The first paper by Que et al. [A14]. The paper presents an FPGA architecture for VAE including a Gaussian random number generator and a layer-wise pipeline architecture. The architecture is implemented by High-Level Synthesis. Compared with CPU and GPU implementations, its latency is reduced by 82x and 208x, respectively. Besides, in the use case of anomaly detection, the proposed VAE architecture is 61x faster than the AE-based architecture, which shows the potential to be applicable to other VAE-based applications.

The second paper by Sun et al. [A15]. This paper modifies the neural network of learned image compression to realize a faster throughput on FPGA. In detail, the number of channels of the neural network is slightly adjusted to ease the routing phase of FPGA architecture. Compared with the previous works, the proposed method achieves an up to 1.5x faster throughput. Besides, its impact on compression efficiency is negligible due to the channel regulation. The proposed algorithm-architecture co-optimization scheme is applicable to other applications as well.

Acknowledgment

We would like to express our gratitude to all the authors who have spent much effort and time in preparing and revising their manuscripts for this Special Issue. We are also in debt to the anonymous reviewers for their insightful feedback on the submissions. Without their exceptional contributions, we would not be able to maintain the high-quality standards for this special issue. We express our gratitude to the Editor-in-Chief (EiC), Prof. Wen-Hsiao Peng, and the Associate EiC, Prof. Shimeng Yu, for their advice and support along the way.

Zhibo Chen, *Corresponding Guest Editor*
School of Information Science and Technology
University of Science and Technology of China
Hefei 230026, China
e-mail: chenzhibo@ustc.edu.cn

Heming Sun, *Guest Editor*
Faculty of Engineering
Yokohama National University
Yokohama 240-0067, Japan
e-mail: sun-heming-vg@ynu.ac.jp

Li Zhang, *Guest Editor*
Multimedia Laboratory
Bytedance Inc.
San Jose, CA 95110 USA
e-mail: lizhang.idm@bytedance.com

Fan Zhang, *Guest Editor*
School of Computer Science
University of Bristol
BS8 1QU Bristol, U.K.
e-mail: fan.zhang@bristol.ac.uk

Appendix: Related Articles

[A1] Z. Chen, H. Sun, L. Zhang, and F. Zhang, “Survey on visual signal coding and processing with generative models: Technologies, standards and optimization,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 2, pp. 149–171, Jun. 2024.

[A2] X. Duan, S. Ma, H. Liu, and C. Jia, “PKU-AIGI-500K: A neural compression benchmark and model for AI-generated images,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 2, pp. 172–184, Jun. 2024.

[A3] Y. Kong, M. Lu, and Z. Ma, “Generative refinement for low bitrate image coding using vector quantized residual,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 2, pp. 185–197, Jun. 2024.

[A4] H. Li and X. Zhang, “Human-machine collaborative image compression method based on implicit neural representations,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 2, pp. 198–208, Jun. 2024.

[A5] P. Du, Y. Liu, and N. Ling, “CGVC-T: Contextual generative video compression with transformers,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 2, pp. 209–223, Jun. 2024.

[A6] C. Sun, H. Yuan, S. Li, X. Lu, and R. Hamzaoui, “Enhancing context models for point cloud geometry compression with context feature residuals and multi-loss,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 2, pp. 224–234, Jun. 2024.

[A7] W. Yan, Y. Sun, G. Yue, W. Zhou, and H. Liu, “FVIFormer: Flow-guided global–local aggregation transformer network for video inpainting,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 2, pp. 235–244, Jun. 2024.

[A8] I. Khalifeh, L. Murn, and E. Izquierdo, “Parameter reduction of kernel-based video frame interpolation methods using multiple encoders,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 2, pp. 245–260, Jun. 2024.

[A9] J. Zhu, V. K. Z. Koh, Z. Lin, and B. Wen, “TM-GAN: A transformer-based multi-modal generative adversarial network for guided depth image super-resolution,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 2, pp. 261–274, Jun. 2024.

[A10] Z. Ma, Y. Wang, H. R. Tohidypour, P. Nasiopoulos, and V. C. M. Leung, “Enhancing image quality by reducing compression artifacts using dynamic window Swin transformer,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 2, pp. 275–285, Jun. 2024.

[A11] Y. Zeng et al., “Physically guided generative adversarial network for holographic 3D content generation from multi-view light field,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 2, pp. 286–298, Jun. 2024.

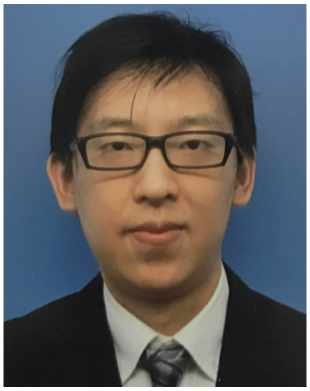
[A12] R. Ji and L. J. Karam, “Compressed-domain vision transformer for image classification,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 2, pp. 299–310, Jun. 2024.

[A13] Z. Wan, H. Qin, R. Xiong, Z. Li, X. Fan, and D. Zhao, “Predicting 360° video saliency: A ConvLSTM encoder–decoder network with spatio-temporal consistency,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 2, pp. 311–322, Jun. 2024.

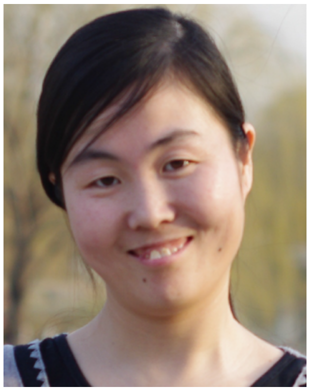
[A14] Z. Que, M. Zhang, H. Fan, H. Li, C. Guo, and W. Luk, “Low latency variational autoencoder on FPGAs,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 2, pp. 323–333, Jun. 2024.

[A15] H. Sun, Q. Yi, and M. Fujita, “FPGA codec system of learned image compression with algorithm-architecture co-optimization,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 2, pp. 334–347, Jun. 2024.

Zhibo Chen (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees from the Department of Electrical Engineering, Tsinghua University, in 1998 and 2003, respectively. He is currently a Full Professor with the University of Science and Technology of China. He has more than 180 publications and over 100 granted patent applications. Some of his standard proposals have been adopted in MPEG/VCEG on video coding and ITU-T P.1202 on video quality assessment. His research interests focus on investigating artificial intelligence technique for advanced visual signal generation, representation, processing and coding, and in other interdisciplinary research fields. He is currently the Chair of the IEEE Visual Signal Processing and Communications Technical Committee (VSPC-TC). He has served as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the Guest Editor for IEEE OPEN JOURNAL OF CIRCUITS AND SYSTEMS (OJCAS) and IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS (JETCAS), the TPC Chair for IEEE PCS 2019, an Organization Committee Member for ICIP 2017 and ICME 2013, and a TPC Member for IEEE ISCAS and VCIP.

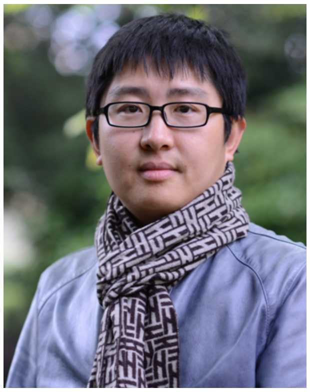


Heming Sun (Member, IEEE) received the B.E. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2011, the double M.E. degrees from Waseda University and Shanghai Jiao Tong University, in 2012 and 2014, respectively, and the Ph.D. degree from Waseda University, in 2017. He was a Researcher at NEC Central Research Laboratories from 2017 to 2018. He was an Assistant Professor at Waseda University from 2018 to 2023. He is currently an Associate Professor with Yokohama National University. His research interests are in algorithms and VLSI architectures for image/video processing and neural networks. He is a member of the Technical Committee on Visual Signal Processing and Communications in the IEEE CAS Society. He got several awards, including the IEEE Computer Society Japan Chapter Young Author Award, the IEEE VCIP Best Paper Award, and the PCS Top-10 Best Paper Award. He is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.



Li Zhang (Senior Member, IEEE) received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2009. She currently leads the Multimedia Laboratory, ByteDance Inc., pioneering cutting-edge technologies in multimedia.

With a focus on video compression, streaming, and signal processing, she holds more than 700 granted U.S. patents and has published more than 100 technical papers in book chapters, journals, and conference proceedings. Additionally, she has made more than 600 adopted standardization contributions to various standards, such as H.266/VVC, H.265/HEVC, AVS, IEEE 1857, H.264/AVC, G-PCC, and JPEG AI. Her research has been recognized with numerous awards, including the Best Paper Award at the 2022 ISCAS Visual Signal Processing and Communications track and the Top 10 Best Paper Award at the 2021 IEEE PCS. She has also secured several first-place accolades in international challenges and received Certificates of Appreciation for her exceptional contributions to the IEEE 1857 standard in 2013 and 2021. She has served as an editor, a software coordinator, and the chair of core experiments in standard groups. She has organized and co-chaired multiple special sessions and grand challenges at conferences. She is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and holds the position of the Publicity Subcommittee Chair of the Technical Committee on Visual Signal Processing and Communications in the IEEE CAS Society.



Fan Zhang (Member, IEEE) received the B.Sc. and M.Sc. degrees from Shanghai Jiao Tong University, Shanghai, China, in 2005 and 2008, respectively, and the Ph.D. degree from the University of Bristol, Bristol, U.K., in 2012. He is currently a Senior Lecturer of visual communications with the School of Computer Science, University of Bristol. He has published over 80 academic articles and has contributed to two books on video compression. His research interests include video coding, video quality assessment, image processing, and deep learning. He was a member of the Organization Committee of PCS 2021. He is a member of the Visual Signal Processing and Communications Technical Committee associated with the IEEE Circuits and Systems Society. He has been serving as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY since 2021.