Guest Editorial
Introduction to the Special Issue on AI-Generated Content for Multimedia

I. Introduction

OUR world is becoming rapidly dependent on data of increasing complexity, diversity, and volume which calls for robust and powerful tools to process such big data. Probabilistic generative models fulfill this goal by learning latent characteristic data relations, especially for the recent emergence of large-scale deep generative models that are able to create realistic content, namely, artificial intelligence-generated content (AIGC). The applications of AIGC span across various domains, and witness rich potential in multimedia content creation, including dialog generation, text-to-speech conversion, image/video generation, and cross-modal content generation.

Despite their success, existing AIGC models still encounter problems such as explainability and expressivity before being applicable in practice, along with ethical concerns regarding misuse and misconduct. More specifically, although being able to generate realistic content, the efficient way to scale up is still yet to be fully understood, as well as the generalization capability to various applications and scenarios. The explainability of AIGC is also of increasing importance by providing clear insights into how decisions and content are generated, to mitigate risks associated with bias, misinformation, and unintended consequences. On the other hand, effective methods to detect generated content are equally important to protect intellectual property between human-created and AI-generated content and maintain trust in digital communications by enabling the verification of content authenticity.

Therefore, this Special Issue focuses on improving AIGC methods for multimedia content, from perspectives of both foundations and applications.

II. Overview of Accepted Papers

After the discussion with the guest editors and reviewers, 20 relevant and high-quality papers from approximately 100 initial submissions are accepted, ranging from the very basics to the cutting-edge applications regarding recent advancements of generative models. We shall briefly review the contributions of the accepted papers, whilst highlighting their novelty that may motivate the follow-up research on AIGC for multimedia scenarios.

A. AIGC Models for Multimedia

Since multimedia data essentially include multiple modalities of content such as voice, images, and videos, there are several papers propose new AIGC models that aim at improving the generation performances across widely used modalities. In [A1], Nazarieh et al. provided the first overview regarding recent state-of-the-art visual content generation methods supported by multimodal information, e.g., text-to-vision and audio-to-vision, together with a comprehensive review of the benchmarking datasets and evaluation metrics. In [A2], Li et al. further proposed a comprehensive database for AI-generated images, which were obtained from extensive prompts and AIGC models. A benchmark experiment on existing widely used metrics was performed against the human subjective scores. For technical contributions, Wang et al. [A3] focused on the text-to-image task and connected the embedding feature from the contrastive language–image pre-training (CLIP) model to the latent generating space of StyleGAN via a proposed mapping network, such that the text-guided image generation is achieved by text-free training style. In [A4], Chen et al. proposed to disentangle identity-relevant information from sample-specific cues by different CLIP encoders, during the finetuning of a pre-trained diffusion model for text-to-image customization and generation. In this way, customization and control over the generated content are enhanced while avoiding common pitfalls associated with overfitting and entangled representations.

B. New Generated Multimedia Content

Moreover, there are also several papers addressing interesting and meaningful generation scenarios. In [A5], Pu et al. proposed new generative neural networks for rule creation and reinforcement learning models for rule assessment. Corresponding metrics were also proposed to control reasonable evolution and adaptation across varying environments, to replicate the dynamics of biological life through computational models. In [A6], Liu et al. proposed to generate talking heads from a single image and one driving audio, namely, one-shot generation, whilst constructing a motion space of video clips to achieve coherent and one-to-many head motion generation. In [A7], Jin et al. focused on generating artworks based on a large-scale generative pre-trained transformer (GPT), in which multi-modal information from user query is aggregated by a dense representation manager as the refined prompt to the GPT. Based on the prompt, multi-task generation is designed for outputting image, video, and music content. In [A8], Zhao et al. focused on generating regional traditional paintings by collecting five types of Chinese folk paintings. The proposed method mimics the human painting process by first introducing multiple content blocks to generate image content while

Digital Object Identifier 10.1109/TCSVT.2024.3427488

preserving the global layout, and then using style blocks to adjust the artistic style, with a focus on lines and edges.

C. Forgery Detection for Generated Content

Along with the development of AIGC methods, the detection of forgery is also crucial for maintaining the integrity of multimedia applications, thus preventing the spread of misguided content. In [A9], Yu et al. proposed to detect the inconsistency between visual and audio signals; this is achieved by comparing deepfake videos with the extracted audio-visual correspondences in real videos via a three-stream network and one cross-modal predictive alignment module. In [A10], Liu et al. further proposed to detect the temporal audio-visual forgery, which can locate the deepfake clips within videos. The proposed method consists of a cross-modal attention mechanism to fuse audio and visual embeddings, supervised by a multi-dimensional contrastive loss to exploit temporal inconsistency. In contrast, Huang et al. [A11] proposed to cheat existing detection methods by leveraging a learning-based notch filtering to remove periodic noise within fake images in an unperceived way, which successfully bypassed existing forgery detection methods. This, on the other hand, also opens a new avenue to improve forgery detection for future research.

D. Multimedia Applications upon AIGC Models

There are also several papers focusing on applications that further extend the scope of sophisticated AIGC models. One gradually agreed-upon application is the usage of AIGC models for image/video compression. Given the AIGC model that generates video frames via sketches, Du et al. [A12] proposed to represent videos by spatio-temporal sketch graphs such that the information is condensed for the follow-up compression. In [A13], Gao et al. further represented visual information into multiple fine-grained domains, including edge maps, regions, and texts, in which the rate-distortion reward function was maximized during training. On the other hand, Gao et al. [A14] developed an invertible image generation module to first transform images into compression-friendly format, such that the images can be compressed at extremely low bit-rates.

Moreover, Wu et al. [A15] proposed a dual transformer model with contrastive learning strategies, which benefits from increased diverse proxy data for the text-based person search task. In [A16], Qin et al. proposed to generate class attention maps as pseudo labels that are able to achieve weakly-supervised semantic segmentation. In [A17], Guan et al. proposed to generate high frequencies within images, and the image super-resolution task was achieved by aggregating low-frequency information obtained from the low-resolution image. In [A18], Liao et al. also applied the generative models for transferring optical images to inverse synthetic aperture radar formats, in which the overall framework was optimized by a meta-learning style. In [A19], Shi et al. combined text-to-image generative models to capture intrinsic semantic relationships between emotion labels, and then to classify object emotions within images. In [A20], Tian et al. further benefited the generative models to achieve noisy image watermark removal in a mixed learning style, i.e., the combination of supervised and self-supervised learning.

III. Conclusion and Acknowledgment

By including new ideas and contributions from the very basics to various applications, this Special Issue essentially responds to the enormous advancements and rapid developments of AIGC methods. We shall also express our gratitude to the authors for their insightful contributions, as well as the reviewers for their valuable comments. We sincerely wish that the methods introduced in this Special Issue could not only help enrich the understanding on the recent progress of AIGC techniques but also pave the way for future research in this exciting and ever-evolving field.

Shengxi Li, *Lead Guest Editor*
School of Electronic and Information Engineering
Beihang University
Beijing 100191, China
e-mail: LiShengxi@buaa.edu.cn

Xuelong Li, *Guest Editor*
School of Artificial Intelligence, OPtics and ElectroNics
Northwestern Polytechnical University
Xi'an 710072, China
e-mail: li@nwpu.edu.cn

Leonardo Chiariglione, *Guest Editor*
CEDEO.net
10040 Villar Dora, Italy
e-mail: leonardo@chiariglione.org

Jiebo Luo, *Guest Editor*
Department of Computer Science
University of Rochester
Rochester, NY 14627 USA
e-mail: jluo@cs.rochester.edu

Wenwu Wang, *Guest Editor*
Department of Electrical and Electronic Engineering
University of Surrey
GU2 7XH Guildford, U.K.
e-mail: w.wang@surrey.ac.uk

Zhengyuan Yang, *Guest Editor*
Microsoft
Redmond, WA 98052 USA
e-mail: zhengyuan.yang13@gmail.com

Danilo Mandic, *Guest Editor*
Department of Electrical and Electronic Engineering
Imperial College London
SW7 2AZ London, U.K.
e-mail: d.mandic@imperial.ac.uk

Hamido Fujita, *Guest Editor*
Department of Software and Information Science
Iwate Prefectural University
Takizawa, Iwate 020-0611, Japan
e-mail: h.fujita@hutech.edu.vn

Appendix: Related Articles

[A1] F. Nazarieh, Z. Feng, M. Awais, W. Wang, and J. Kittler, "A survey of cross-modal visual content generation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6814–6832, Aug. 2024, doi: 10.1109/TCSVT.2024.3351601.

[A2] C. Li et al., "AGIQA-3K: An open database for AI-generated image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6833–6846, Aug. 2024, doi: 10.1109/TCSVT.2023.3319020.

[A3] Y. Wang, W. Zhou, J. Bao, W. Wang, L. Li, and H. Li, "CLIP2GAN: Towards bridging text with the latent space of GANs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6847–6859, Aug. 2024, doi: 10.1109/TCSVT.2023.3294261.

[A4] H. Chen, Y. Zhang, X. Wang, X. Duan, Y. Zhou, and W. Zhu, "DisenDreamer: Subject-driven text-to-image generation with sample-aware disentangled tuning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6860–6873, Aug. 2024, doi: 10.1109/TCSVT.2024.3369757.

[A5] J. Pu, H. Duan, J. Zhao, and Y. Long, "Rules for expectation: Learning to generate rules via social environment modelling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6874–6887, Aug. 2024, doi: 10.1109/TCSVT.2023.3334526.

[A6] J. Liu et al., "OSM-Net: One-to-many one-shot talking head generation with spontaneous head motions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6888–6900, Aug. 2024, doi: 10.1109/TCSVT.2024.3394850.

[A7] C. Jin, R. Zhu, Z. Zhu, L. Yang, M. Yang, and J. Luo, "MtArtGPT: A multi-task art generation system with pre-trained transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6901–6912, Aug. 2024, doi: 10.1109/TCSVT.2024.3349567.

[A8] Y. Zhao, H. Li, Z. Zhang, Y. Chen, Q. Liu, and X. Zhang, "Regional traditional painting generation based on controllable disentanglement model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6913–6925, Aug. 2024, doi: 10.1109/TCSVT.2023.3298811.

[A9] Y. Yu, X. Liu, R. Ni, S. Yang, Y. Zhao, and A. C. Kot, "PVASS-MDD: Predictive visual-audio alignment self-supervision for multimodal deepfake detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6926–6936, Aug. 2024, doi: 10.1109/TCSVT.2023.3309899.

[A10] M. Liu, J. Wang, X. Qian, and H. Li, "Audio-visual temporal forgery detection using embedding-level fusion and multi-dimensional contrastive loss," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6937–6948, Aug. 2024, doi: 10.1109/TCSVT.2023.3326694.

[A11] Y. Huang, F. Juefei-Xu, Q. Guo, Y. Liu, and G. Pu, "Dodging DeepFake detection via implicit spatial-domain notch filtering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6949–6962, Aug. 2024, doi: 10.1109/TCSVT.2023.3325427.

[A12] Q. Du, Y. Duan, Z. Xie, X. Tao, L. Shi, and Z. Jin, "Optical flow-based spatiotemporal sketch for video representation: A novel framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6963–6977, Aug. 2024, doi: 10.1109/TCSVT.2023.3349130.

[A13] J. Gao, C. Jia, Z. Huang, S. Wang, S. Ma, and W. Gao, "Rate-distortion optimized cross modal compression with multiple domains," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6978–6992, Aug. 2024, doi: 10.1109/TCSVT.2024.3364153.

[A14] F. Gao, X. Deng, J. Jing, X. Zou, and M. Xu, "Extremely low bit-rate image compression via invertible image generation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6993–7004, Aug. 2024, doi: 10.1109/TCSVT.2023.3317424.

[A15] H. Wu, W. Chen, Z. Liu, T. Chen, Z. Chen, and L. Lin, "Contrastive transformer learning with proximity data generation for text-based person search," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7005–7016, Aug. 2024, doi: 10.1109/TCSVT.2023.3329220.

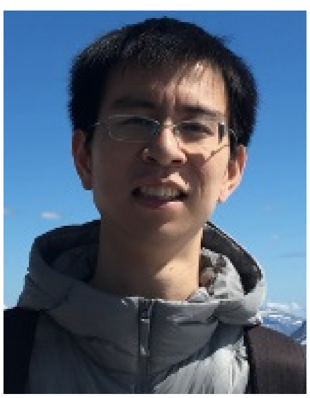
[A16] Z. Qin et al., "Enhanced pseudo-label generation with self-supervised training for weakly-supervised semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7017–7028, Aug. 2024, doi: 10.1109/TCSVT.2024.3364764.

[A17] W. Guan, H. Li, D. Xu, J. Liu, S. Gong, and J. Liu, "Frequency generation for real-world image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7029–7040, Aug. 2024, doi: 10.1109/TCSVT.2024.3367876.

[A18] H. Liao, J. Xia, Z. Yang, F. Pan, Z. Liu, and Y. Liu, "Meta-learning based domain prior with application to optical-ISAR image translation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7041–7056, Aug. 2024, doi: 10.1109/TCSVT.2023.3318401.

[A19] G. Shi, S. Deng, B. Wang, C. Feng, Y. Zhuang, and X. Wang, "One for all: A unified generative framework for image emotion classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7057–7068, Aug. 2024, doi: 10.1109/TCSVT.2023.3341840.

[A20] C. Tian, M. Zheng, B. Li, Y. Zhang, S. Zhang, and D. Zhang, "Perceptive self-supervised learning network for noisy image watermark removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7069–7079, Aug. 2024, doi: 10.1109/TCSVT.2024.3349678.



Shengxi Li (Member, IEEE) received the Ph.D. degree in electrical and electronic engineering from Imperial College London, London, U.K., in 2021. He is currently a Professor with the School of Electronic and Information Engineering, Beihang University, Beijing, China. His research interests include generative models, statistical signal processing, and machine learning. He was a recipient of the Young Investigator Award of the International Neural Network Society.

Xuelong Li (Fellow, IEEE) has been the Chief Technology Officer (CTO) of China Telecom since 2023. Before that, he was a Full Professor with Northwestern Polytechnical University, Xi'an, China, from 2018 to 2023, a Full Professor with the Chinese Academy of Sciences from 2009 to 2018, a Lecturer/Senior Lecturer/Reader with the University of London from 2004 to 2009, a Lecturer with Ulster University from 2003 to 2004, and he previously took positions at the Chinese University of Hong Kong, The University of Hong Kong, Microsoft Research, and Huawei Technologies Company Ltd. He is currently a Full Professor with the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University.

Leonardo Chiariglione received the M.S. degree from the Polytechnic of Turin, Turin, Italy, and the Ph.D. degree from The University of Tokyo, Tokyo, Japan. He founded and chaired the Moving Picture Experts Group (MPEG) for 32 years and proposed and chairs the Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Organization, for AI-enabled data coding standards. He was a recipient of several awards, among which are: the IBC John Tucker Award, the IEEE Masaru Ibuka Consumer Electronics Award, and the Kilby Foundation Award.

Jiebo Luo (Fellow, IEEE) is currently the Albert Arendt Hopeman Professor of engineering and a Professor of computer science with the University of Rochester, Rochester, NY, USA, which he joined in 2011 after a prolific career of 15 years with Kodak Research Laboratories. He has authored over 600 technical articles and holds more than 90 U.S. patents. His research interests include computer vision, natural language processing (NLP), machine learning, data mining, computational social science, and digital health. He is a fellow of the U.S. National Academy of Inventors (NAI), the Association for Computing Machinery (ACM), the Association for the Advancement of Artificial Intelligence (AAAI), the Society of Photographic Instrumentation Engineers (SPIE), and the International Association for Pattern Recognition (IAPR). He has been involved in numerous technical conferences, including serving as the Program Co-Chair for ACM Multimedia 2010, the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012, the ACM International Conference on Multimedia Retrieval (ICMR) 2016, and the IEEE International Conference on Image Processing (ICIP) 2017, and the General Co-Chair for ACM Multimedia 2018 and the IEEE International Conference on Multimedia & Expo (ICME) 2024. He has served on the editorial boards for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON BIG DATA, *ACM Transactions on Intelligent Systems and Technology* (TIST), *Pattern Recognition*, *Machine Vision and Applications*, *Health Data Science*, and *Intelligent Medicine*. He was the Editor-in-Chief of IEEE TRANSACTIONS ON MULTIMEDIA from 2020 to 2022.

Wenwu Wang (Senior Member, IEEE) was born in Anhui, China. He received the B.Sc., M.E., and Ph.D. degrees from Harbin Engineering University, China, in 1997, 2000, and 2002, respectively. He was with King's College London, London, U.K.; Cardiff University, Cardiff, U.K.; Tao Group Ltd. (now Antix Labs Ltd.); and Creative Labs, before joining the University of Surrey, Guildford, U.K., in May 2007. He is currently a Professor of signal processing and machine learning with the University of Surrey. He is also an AI Fellow with Surrey Institute for People-Centred Artificial Intelligence. He has coauthored more than 300 articles in these areas. His current research interests include signal processing, machine learning and perception, artificial intelligence, machine audition (listening), and statistical anomaly detection. During 2021–2026, he is an elected member of the IEEE SPS Signal Processing Theory and Methods Technical Committee. During 2023–2024, he is a Board Member of the IEEE SPS Technical Directions Board. He has been recognized as the co-author or a co-recipient of more than 15 awards, including the 2022 IEEE Signal Processing Society Young Author Best Paper Award, the ICAUS 2021 Best Paper Award, the DCASE 2020 and 2023 Judge's Award, the DCASE 2019 and 2020 Reproducible System Award, and the LVA/ICA 2018 Best Student Paper Award. During 2023–2024, he is the elected Chair of the IEEE Signal Processing Society (SPS) Machine Learning for Signal Processing Technical Committee. During 2022–2024, he is the Vice Chair of the EURASIP Technical Area Committee on Acoustic Speech and Music Signal Processing, where he is also the Chair during 2025–2027. He was the Satellite Workshop Co-Chair of INTERSPEECH 2022, the Publication Co-Chair of IEEE ICASSP 2019, the Local Arrangement Co-Chair of IEEE MLSP 2013, and the Publicity Co-Chair of IEEE SSP 2009. He is the Satellite Workshop Co-Chair of IEEE ICASSP 2024, the Special Session Co-Chair of IEEE MLSP 2024, and the Technical Program Co-Chair of IEEE MLSP 2025. He is an Associate Editor of IEEE/ACM TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING during 2020–2025 and an Associate Editor for IEEE TRANSACTIONS ON MULTIMEDIA during 2024–2026. From 2014 to 2018, he was an Associate Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING. From 2019 to 2023, he was the Senior Area Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING.

Zhengyuan Yang received the B.E. degree in electrical engineering from the University of Science and Technology of China (USTC) in 2016 and the Ph.D. degree in computer science from the University of Rochester (UR) in 2021. He is currently a Senior Researcher with Microsoft. His research interests include the intersection of computer vision and natural language processing, including multi-modal vision-language understanding and generation. He was a recipient of the ACM SIGMM Award for Outstanding Ph.D. Thesis, the Twitch Research Fellowship, and the ICPR 2018 Best Industry Related Paper Award.

Danilo Mandic (Fellow, IEEE) received the Ph.D. degree in nonlinear adaptive signal processing from Imperial College London, London, U.K., in 1999.

He is currently a Professor of machine intelligence with Imperial College London. His publication record includes two research monographs titled *Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability* (Wiley, First Edition, August 2001) and *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models* (Wiley, First Edition, April 2009), an edited book titled *Signal Processing Techniques for Knowledge Extraction and Information Fusion* (Springer, 2008), and a two-volume research monograph titled *Tensor Networks for Dimensionality Reduction and Large Scale Optimization* (Now Publishers, 2016 and 2017). He received the Denis Gabor Award for Outstanding Achievements in Neural Engineering by the International Neural Networks Society, the 2018 Best Paper Award from *IEEE Signal Processing Magazine* for his article on tensor decompositions, and the President Award for Excellence in Postgraduate Supervision at Imperial College London. He has been an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS, IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS.

Hamido Fujita (Life Senior Member, IEEE) received the Ph.D. degree (Hons.) from Óbuda University, Budapest, Hungary, in 2013, and the Ph.D. degree (Hons.) from Timisoara Technical University, Timisoara, Romania, in 2018. He received the title of an Honorary Professor from Óbuda University in 2011. He is currently a Professor with Malaysia–Japan International Institute of Technology, Universiti Teknologi Malaysia, and a Distinguished Professor with Iwate Prefectural University, Japan. He is also a Research Professor with the University of Granada, Spain. He is a Distinguished Research Professor with the University of Granada and an Adjunct Professor with Taipei Technical University, Taiwan; Harbin Engineering University, China; and others. He supervised Ph.D. students jointly with the University of Laval, Quebec City, QC, Canada; the University of Technology Sydney; Oregon State University, Corvallis, OR, USA; the University of Paris 1 Pantheon-Sorbonne, Paris, France; and the University of Genoa, Italy. He headed a number of projects, including intelligent HCI, a project related to mental cloning for healthcare systems as an intelligent user interface between human-users and computers, and SCOPE project on virtual doctor systems for medical applications. He collaborated with several research projects in Europe, and recently he is collaborating in OLIMPIA project supported by Tuscany region on therapeutic monitoring of Parkinson disease. He has published more than 400 articles mainly in high impact factor journals. He was a recipient of the Honorary Scholar Award from the University of Technology Sydney in 2012. He was the Editor-in-Chief of *Knowledge-Based Systems* (Elsevier) (2005–2019). Since 2020, he has been an Emeritus Editor of *Knowledge-Based Systems*. Since 2020, he has been the Editor-in-Chief of *Applied Intelligence* (Springer) and the Editor-in-Chief of *International Journal of Healthcare Management* (Taylor & Francis). He is the Executive Chairperson of the i-SOMET Incorporated Association, Japan. He was a Highly Cited Researcher in Cross-Field for the year 2019 and in computer science by Clarivate Analytics for the year 2020, 2021, 2022, and 2023, respectively.