Guest Editorial Introduction to the Special Issue on Video Transformers

I. Introduction

CURRENTLY, Transformer has been widely used in natural language and image processing and has achieved excellent results. Benefiting from the self-attention operation and global interaction, Transformer has demonstrated more powerful spatiotemporal modeling capabilities than traditional convolutional and recurrent neural networks. However, research on video Transformer is still in its infancy. Specifically, with the development of internet technology, video data has become a commonly used medium, playing a critical role in many areas such as entertainment, education, healthcare, security, etc. Different from static data such as images and text, video data consists of a series of image frames and is more concerned with temporal and motion information, which makes it necessary to employ some adaptations and well-designed network architectures to capture the discriminative features. In addition, the multi-modal information attached to video data further increases the difficulty of applying Transformer to videos.

Therefore, this special section focuses on exploring video Transformer, with the aim of bringing new and insightful enlightenment to researchers working on various video tasks and providing effective solutions to fix frontier video-related issues.

II. Overview of Accepted Articles

This special section covers scenarios where a video Transformer has been applied to different video-related tasks, including classification, generation, object segmentation and tracking, and human pose estimation and motion prediction. In total, 12 relevant and high-quality articles are selected by the guest editors, including three articles on classification tasks, three articles on generation tasks, four articles on object segmentation and tracking, and two articles on human pose estimation and motion prediction. In the following section, we will briefly summarize each article, highlighting their contributions and innovations.

A. Classification Tasks

Classification is a fundamental and crucial task in artificial intelligence, with numerous applications in various domains. In the medical field, endoscopic image classification for the digestive tract is particularly important. By leveraging deep learning models to classify multiple frames captured by endoscopy, doctors can rapidly and accurately diagnose the condition of patients. Wang et al. [A1] propose a novel vision Transformer model based on hybrid shifted windows, which can well capture both short-range and long-range dependency and alleviate the limitation of Swim Transformer in capturing long-range dependence in complex gastrointestinal endoscopy images. With the ongoing advancement of DeepFake technology, DeepFake video detection also draws the attention of researchers. By classifying videos as true or false, the authenticity of videos can be effectively discerned, thus avoiding misinformation and deception. Existing DeepFake video detection methods fail to obtain fine-grained spatiotemporal information, leading to limited generalization ability. Yu et al. [A2] design a novel multiple spatiotemporal views Transformer (MSVT) using the local and global spatiotemporal views to mine subtle and comprehensive spatiotemporal clues for generalized DeepFake detection. A novel global-local structure is proposed to effectively integrate multilevel features, thus achieving excellent improvements. In the realm of action tasks, early action prediction also has garnered significant attention. This task, akin to action recognition, involves classifying patterns of action. However, the key distinction lies in that early action prediction seeks to identify the action as expeditiously as possible before it is fully executed. Guan et al. [A3] present a multimodal Transformer-based dual action prediction model which contains two key modules: the early segment action prediction module and the future segment action prediction module. A consistency regularizer is introduced to mine the coherence of the full video segment. They also design a two-stage optimization scheme, including the mutual enhancement stage and end-to-end aggregation stage, to make the most of the two modules.

B. Generation Tasks

The generation task refers to generating some new data through algorithms or models, which can be images, audio, video, text, etc. Among these, generating video captions is a very meaningful task. By analyzing and processing the video, captions for the video can be generated automatically, which brings great convenience to people to understand the video content better. Wu et al. [A4] propose a unified framework CAT for video captioning. Two modules called concept parser and multi-modal graph are presented to extract

Digital Object Identifier 10.1109/TCSVT.2023.3294789

high-level cues and bridge the gaps between different levels of representations, respectively. Benefiting from the proposed modules, CAT generates a better caption than the previous methods. In addition, video super-resolution is a task of great interest. The aim of the task is to convert low-resolution video to high-resolution video, thereby improving the quality and clarity of the video. Zhang et al. [A5] first come up with a new bicubic up-sampling method for multiscale acquisition. Then convolution and self-attention techniques are utilized to correct and align features at different scales. In addition, they introduce a novel approach to constructing the activation function, which effectively solves the typical issues of complex calculation or low continuity of existing activation functions. There are also some fulfilling and rewarding generative tasks such as lip reading. The lip reading task automatically generates human speech content by analyzing the shape of the lips, thus helping people with hearing impairments to better understand what others are communicating. Xue et al. [A6] propose a cross-modal Transformer framework for sentence-level lipreading, which is capable of generalizing to unseen speakers by utilizing landmarks as motion trajectories to caliber the visual variations. The model improves the alignment of heterogeneous features by cross-modal fusion suggested by cross-attention.

C. Object Segmentation and Tracking Tasks

The object segmentation task is to separate the object of interest from the background in an image or video, while the tracking task refers to tracking the position of the object in a video. Both contribute to the identification and analysis of targets throughout the video sequence. Referring video object segmentation (RVOS) is a special segmentation task that aims to segment text-described objects from video sequences. Gao et al. [A7] present the decoupled multimodal Transformer (DMFormer) for RVOS, which facilitates explicit interaction between visual features and different syntactic components in text. They also propose an effective strategy to transfer knowledge from large-scale pretrained vision-language alignment to RVOS for better performance. Similar to RVOS, natural language tracking is a special kind of tracking task that aims to localize the text-described object using a sequence of bounding boxes in video frames. Wang et al. [A8] propose an end-to-end unified network based on Transformer with a novel selective feature gather module, which employs two pairs of encoder and decoder structures for collaboratively grounding and tracking learning. It can not only conduct grounding and tracking learning independently but also tackle the natural language tracking task by integrating the information with varying semantics, resulting in improved accuracy and greater robustness. Zheng et al. [A9] introduce a new TaTrack model for visual single-object tracking that merges the Transformer and siamese neural network. Benefiting from the proposed target-aware module and the update strategy, TaTrack boosts the global interaction ability of the model. For multidrone tracking, Chen et al. [A10] design a Transformer-based collaborative single-object tracking framework TransMDOT to automatically model the association between multiple templates and search regions. TransMDOT consists of a multidrone relation modeling module, a cross-drone mapping module, and a system perception fusion module and is effective in handling information interaction among drones. A new evaluation metric SPFI is also proposed to comprehensively evaluate the state of the tracker during the system fusion period.

D. Human Pose Estimation and Motion Prediction Tasks

For exploring the human body, deep learning has made great strides in two important research directions, including human pose estimation and human motion prediction. Human pose estimation identifies key point locations and pose information of the human body, which can be applied in areas such as human–computer interaction and virtual reality. Human motion prediction can predict future movements from historical motion information for applications such as motion analysis and health management. Gai et al. [A11] present an effective Transformer-based framework SLT-Pose for human pose estimation, which contains four modules, including a personalized feature extraction module, self-feature refinement module, cross-frame temporal learning module, and disentangled keypoint detector. Extensive experiments show that SLT-Pose outperforms state-of-the-art methods in both objective evaluation and subjective visual performance. For human motion prediction, Chen et al. [A12] design a novel motion characterization method based on instantaneous momentum from the perspective of Newtonian mechanics. To effectively capture both local and global temporal patterns inherent in motion sequences, they introduce a dual-stream architecture and a TA-GCN module. The proposed network is capable of accurately modeling the spatiotemporal coherence of the human body, resulting in improved prediction performance.

III. Conclusion and Acknowledgments

In general, we hope that these articles will bring inspiration to researchers and further drive the progress and development of the entire field of video understanding. We would like to express our sincere gratitude to the authors of the selected articles for their contributions to this special section. We also call on more researchers to explore various applications of Transformers for more possibilities in the field of video. As we look to the future, video Transformers have the potential to become a critical tool in video processing, providing powerful support in understanding and analyzing video content.

Liqiang Nie, *Lead Guest Editor*
School of Computer Science and Technology
Harbin Institute of Technology (Shenzhen)
Shenzhen 518055, China
e-mail: nieliqiang@gmail.com

Jianlong Wu, *Guest Editor*
School of Computer Science and Technology
Harbin Institute of Technology (Shenzhen)
Shenzhen 518055, China
e-mail: wujianlong@hit.edu.cn

NICU SEBE, *Guest Editor*
Department of Information Engineering and Computer Science
University of Trento
38122 Trento, Italy
e-mail: niculae.sebe@unitn.it

KIYOHARU AIZAWA, *Guest Editor*
Department of Information and Communication Engineering
University of Tokyo
Tokyo 113-8654, Japan
e-mail: aizawa@hal.t.u-tokyo.ac.jp

APPENDIX: RELATED ARTICLES

[A1] W. Wang, X. Yang, and J. Tang, "Vision transformer with hybrid shifted windows for gastrointestinal endoscopy image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4452–4461, Sep. 2023.

[A2] Y. Yu et al., "MSVT: Multiple spatiotemporal views transformer for DeepFake video detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4462–4471, Sep. 2023.

[A3] W. Guan et al., "Egocentric early action prediction via multimodal transformer-based dual action prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4472–4483, Sep. 2023.

[A4] B. Wu, B. Liu, P. Huang, J. Bao, X. Peng, and J. Yu, "Concept parser with multi-modal graph learning for video captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4484–4495, Sep. 2023.

[A5] F. Zhang, G. Chen, H. Wang, J. Li, and C. Zhang, "Multi-scale video super-resolution transformer with polynomial approximation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4496–4506, Sep. 2023.

[A6] F. Xue, Y. Li, D. Liu, Y. Xie, L. Wu, and R. Hong, "LipFormer: Learning to lipread unseen speakers based on visual-landmark transformers," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4507–4517, Sep. 2023.

[A7] M. Gao, J. Yang, J. Han, K. Lu, F. Zheng, and G. Montana, "Decoupling multimodal transformers for referring video object segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4518–4528, Sep. 2023.

[A8] R. Wang et al., "Unified transformer with isomorphic branches for natural language tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4529–4541, Sep. 2023.

[A9] Y. Zheng, Y. Zhang, and B. Xiao, "Target-aware transformer tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4542–4551, Sep. 2023.

[A10] G. Chen, P. Zhu, B. Cao, X. Wang, and Q. Hu, "Cross-drone transformer network for robust single object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4552–4563, Sep. 2023.

[A11] D. Gai et al., "Spatiotemporal learning transformer for video-based human pose estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4564–4576, Sep. 2023.

[A12] H. Chen, J. Hu, W. Zhang, and P. Su, "Spatiotemporal consistency learning from momentum cues for human motion prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4577–4587, Sep. 2023.

Liqiang Nie (Senior Member, IEEE) received the B.Eng. degree from Xi'an Jiaotong University and the Ph.D. degree from the National University of Singapore (NUS). After the Ph.D. degree, he continued his research with NUS, as a Research Fellow for three years. He is currently a Professor with the Harbin Institute of Technology (Shenzhen). He has coauthored more than 200 articles and four books. He received more than 20 000 Google Scholar citations as of July 2023. His research interests include multimedia computing and information retrieval. He received many awards, such as ACM MM and SIGIR Best Paper Honorable Mention in 2019, the SIGMM Rising Star in 2020, the TR35 China 2020, the DAMO Academy Young Fellow in 2020, the SIGIR Best Student Paper in 2021, and the ACM MM Best Paper in 2022. He is the Area Chair of ACM MM 2018–2023. He is an AE of IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON MULTIMEDIA, *ACM ToMM*, and *Information Sciences*.

Jianlong Wu (Member, IEEE) received the B.Eng. degree from the Huazhong University of Science and Technology in 2014 and the Ph.D. degree from Peking University in 2019. He is currently an Associate Professor with the Harbin Institute of Technology (Shenzhen). He has published more than 20 papers in top journals and conferences, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, ICML, NeurIPS, and ICCV. His research interests include computer vision and machine learning, especially unsupervised and semi-supervised learning. He serves as a Senior Program Committee Member for IJCAI 2021, the Area Chair for ICPR 2022/2020, and a Reviewer for many top journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *IJCV*, ICML, and CVPR. He received many awards, such as the Outstanding Reviewer of ICML 2020 and the Best Student Paper of SIGIR 2021.

Nicu Sebe (Senior Member, IEEE) is currently a Professor in computer science with the University of Trento, Trento, Italy. He is leading the research in the areas of multimedia information retrieval and human–computer interaction in computer vision applications. He is a fellow of IAPR and a Senior Member of ACM. He was involved in the organization of the major conferences and workshops addressing the computer vision and human-centered aspects of multimedia information retrieval, among which as the General Co-Chair of the IEEE Automatic Face and Gesture Recognition Conference (FG 2008) and the ACM International Conference on Image and Video Retrieval (CIVR) in 2007 and 2010. He was the General Chair of the ACM Multimedia 2013 and ACM ICMR 2017 and the Program Chair of ACM Multimedia 2011 and 2007, ECCV 2016, and ICCV 2017. He was the Program Chair of ICPR 2020 and the ACM SIGMM Vice Chair.

Kiyoharu Aizawa (Fellow, IEEE) received the B.E., M.E., and Dr.Eng. degrees in electrical engineering from The University of Tokyo, in 1983, 1985, and 1988, respectively. He is currently a Professor at the Department of Information and Communication Engineering, The University of Tokyo. He was a Visiting Assistant Professor at the University of Illinois from 1990 to 1992. His research interests include multimedia applications, image processing, and computer vision. He is a fellow of IEICE and ITE and a Council Member of the Science Council of Japan. He received the 1987 Young Engineer Award, the 1990 and 1998 Best Paper Awards, the 1991 Achievement Award, the 1999 Electronics Society Award from IEICE Japan, the 1998 Fujio Frontier Award, the 2002 and 2009 Best Paper Award, and the 2013 and 2020 Achievement Award from ITE Japan. He received the IBM Japan Science Prize in 2002. He serves as an Associate Editor for ACM TOMM and served as the Editor-in-Chief for *Journal of ITE Japan* and an Associate Editor of IEEE MULTIMEDIA, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE TRANSACTIONS ON MULTIMEDIA. He was the President of ITE and ISS Society of IEICE, in 2019 and 2018, respectively. He has served a number of international and domestic conferences. He was the General Co-Chair of ACM Multimedia 2012 and ACM ICMR2018.