Machine Intelligence Research
www.mi-research.net

21(4), August 2024, 615-616
DOI: 10.1007/s11633-024-1513-5

# Editorial for Special Issue on Multi-modal Representation Learning

The past decade has witnessed the impressive and steady development of single-modal AI technologies in several fields, thanks to the emergence of deep learning. Less studied, however, is multi-modal AI – commonly considered the next generation of AI – which utilizes complementary context concealed in different-modality inputs to improve performance. Humans naturally learn to form a global concept from multiple modalities (i.e., sight, hearing, touch, smell, and taste), even when some are incomplete or missing. Thus, in addition to the two popular modalities (vision and language), other types of data such as depth, infrared information, and events are also important for multi-modal learning in real-world scenes.

This special issue is to bring smart solutions together for potentially robust representation learning in multi-modal scenes. We are interested in works related to theoretical, algorithmic, metric, and dataset advances, as well as new applications. Finally, seven papers were accepted to form this special issue from the perspective of improving performance in visual or language modalities. Although these works are not directly based on multi-modal representation learning, we still believe that these works will also bring more potential value to improving multi-modal representation learning.

The first paper, entitled "Segment Anything Is Not Always Perfect: An Investigation of SAM on Different Real-world Applications" from Wei Ji et al., investigates large vision model-segmentation anything model (SAM)′s performance across various applications, e.g., natural image, agriculture, manufacturing, remote sensing, and healthcare. The authors conduct an analysis and discussion of the advantages and constraints of SAM, and they also provide a perspective on its future evolution in generic segmentation tasks.

The second paper, entitled "Rethinking Polyp Segmentation from an Out-of-Distribution Perspective" from Ge-Peng Ji et al., contributes to medical image segmentation task from an out-of-distribution perspective. They leverage the ability of masked autoencoders to learn in-distribution representations, then utilize reconstruction and inference, with feature space standardization to align the latent distribution of the diverse abnormal samples with the statistics of the healthy samples. This method is quite different from previous fully-supervised polyp segmentation (PS) approaches, and achieves very good results on six PS benchmarks

The third paper, entitled "Rethinking Global Context in Crowd Counting" by Guolei Sun et al., focuses on the topic of crowd counting from the views of global context. The authors introduce a context token to the transformer input sequence in the token-attention module (TAM) and regression-token module (RTM) so that it can recalibrate encoded features and predict the total person count of the image. This solution will inspire future multimodal interactions.

The fourth paper, entitled "Towards Domain-agnostic Depth Completion" from Guangkai Xu et al., explores a robust and simple system for depth completion. First, they leverage a single image depth prior and their model is trained with diverse data augmentation. Then, two new depth completion benchmarks have been designed to assess the model′s generalization ability from the perspective of robustness to noise, different sparsity patterns, and diverse scenes. This scheme provides an idea to high-quality mobile depth capture, which can improve downstream depth-related applications on mobile devices.

The fifth paper, entitled "Vision Transformers with Hierarchical Attention" from Yun Liu et al, addresses the inefficiency inherent in vanilla vision transformers due to the elevated computational and space complexity associated with MHSA. To alleviate the computational and space demands, they introduce a novel hierarchical component H-MHSA for MHSA computation. H-MHSA can directly capture both global dependencies and local relationships. By introducing H-MHSA into the network, they formulate the HAT-Net family, showcasing its prowess in advancing vision representation learning.

The sixth article is a method of CNN architecture, titled as "A Novel Divide and Conquer Solution for Long-term Video Salient Object Detection" by Yun-Xiao Li et al, to solve complex pattern modeling in long video. They design a divide-and-conquer framework, which can convert a complex problem domain into multiple simple ones. The key is a novel background consistency analysis (BCA) which effectively divides the mined frames into disjoint groups. For each group, the authors assign an individual deep model to capture its key attribute during the fine-tuning phase. During the testing stage, they fur-

Editorial
Special Issue on Multi-modal Representation Learning
Colored figures are available in the online version at https://link.springer.com/journal/11633

ther use a model-matching strategy, which could dynamically select the best-matched model from those fine-tuned ones to handle the given testing frame.

The last paper, entitled "TextFormer: A Query-based End-to-end Text Spotter with Mixed Supervision" by Yukun Zhai et al, exploits the complementary nature of text detection and recognition in a simple yet robust query-based end-to-end text spotting framework called TextFormer. It prevents RoI operation and the complex post-processing procedure. By using the encoder-decoder module to compute shared semantic features for each text query, the following multitask branches can be co-optimized from the semantic features. Additionally, they develop an AGG module to extract global features for recognizing arbitrarily shaped texts. Further, a mixed-supervision learning manner is employed to train the model with different levels of annotations to tackle the inconsistency between text detection and recognition. It is worth mentioning that the authors show that text detection and recognition can mutually improve each other, especially for ambiguous texts.

Among the seven articles including six technical and one viewpoint papers; three are from Chinese institutions and four are from institutions of Canada, Switzerland, the United Kingdom, and Singapore. We believe that with the active exploration of researchers, multi-modal representation learning topics will play a positive role in the publication of this special issue.

*Deng-Ping Fan*
*Nankai University, China & ETH Zürich, Switzerland*
*Nick Barnes*
*Australian National University, Australia*
*Ming-Ming Cheng*
*Nankai University, China*
*Luc Van Gool*
*ETH Zürich, Switzerland*

**Deng-Ping Fan** received the Ph.D. degree from Nankai University, China in 2019. He joined ETH Zürich, Switzerland in 2022. Currently, he is a professor at the College of Computer Science in Nankai University, China. He has published about 30 top journal and conference papers such as TPAMI, CVPR, ICCV, ECCV, etc. He won the Best Paper Finalist Award at IEEE CVPR 2019, and the Best Paper Award Nominee at IEEE CVPR 2020.

His research interests include computer vision and visual attention, especially in RGB salient object detection (SOD), RGBD SOD, Video SOD, and CoSOD.

E-mail: dengpfan@gmail.com (Corresponding author)
ORCID iD: 0000-0002-5245-7518

**Nick Barnes** received the B.Sc. (Hons) degree and Ph.D. degree in engineering in robotic vision from the University of Melbourne, Australia in 1994 and 1999, respectively. He is currently a professor with the School of Computing, Australian National University (ANU), Australia. He was a visiting researcher with the LIRA Lab at the University of Genoa, Italy in 1999, and a tenured lecturer with the University of Melbourne, Australia until 2003. He was then with NICTA, an ICT Centre of Excellence, Australia from 2003–2016, where he was a senior principal researcher and was executive leader of the Computer Vision Research Group. He was with CSIRO from 2016–2019, where he led the Computer Vision Research Group. He has best paper awards and nominations including from CVPR, Robotics and Systems Science, IROS, MICCAI Wshp on Computer Assisted Endscopy, DICTA. He has multiple patents in the area of vision processing for prosthetic vision, which contributed the creation of the company Bionic Vision Technologies, Australia.

His research interests include weakly supervised dense prediction, 3D vision, and computer vision for prosthetic vision.

E-mail: nick.barnes@anu.edu.au
ORCID iD: 0000-0002-9343-9535

**Ming-Ming Cheng** received the Ph.D. degree in computer science and technology from Tsinghua University, China in 2012. Then, he did two years research fellow with Professor Philip Torr in Oxford, UK. He is now a professor at Nankai University, China, leading the Media Computing Laboratory. He received research awards, including ACM China Rising Star Award, IBM Global SUR Award, and CCF-Intel Young Faculty Researcher Program. He is on the editorial boards of IEEE TPAMI/TIP, MIR.

His research interests include computer graphics, computer vision and image processing.

E-mail: cmm@nankai.edu.cn
ORCID iD: 0000-0001-5550-8758

**Luc Van Gool** received the B.Eng. degree in electromechanical engineering from the Katholieke Universiteit Leuven, Belgium in 1981. Currently, he is a professor at the Katholieke Universiteit Leuven, Belgium, and the ETH Zürich, Switzerland. He leads computer vision research at both places and also teaches at both. He has been a program committee member of several major computer vision conferences. He received several Best Paper Awards, won a David Marr Prize and a Koenderink Award, and was nominated Distinguished Researcher by the IEEE Computer Science Committee. He is a co-founder of 10 spin-off companies.

His research interests include 3D reconstruction and modeling, object recognition, tracking, gesture analysis, and a combination of those.

E-mail: vangool@vision.ee.ethz.ch
ORCID iD: 0000-0002-3445-5711