Guest Editorial
Learning From Noisy Multimedia Data

THE MULTIMEDIA age and its proliferation of devices and platforms is fueling exponential data growth. As computational power and deep learning algorithms rapidly evolve, the web has become a rich source of potential training data for robust machine learning, with search engines such as Google and Bing, Twitter, TikTok, Instagram, and short video sharing platforms offering large-scale data points in the hundreds of millions. The concurrent shift in the Internet to richer web data modalities such as text, audio, image, and video reveal further opportunities to leverage large-scale data for the automatic construction of a variety of datasets for model training and testing. However, the ubiquity of multimedia data means noise is a fundamental challenge, with 'label noise' and 'domain mismatch' the most critical issues in automatically collected datasets. Learning from noisy multimedia data tends towards poor performance, making it increasingly essential to address these challenges.

This special issue provides a premier forum for researchers in multimedia big data to share challenges and recent advancements in learning from noisy multimedia data. The growing interest in this area of multimedia research is evidenced by the large volume of submissions to this issue, with a total of 72 papers received. This necessitated a two-round review process for the guest editorial team, with 38 of the 72 papers selected in the first-round evaluation as closely matching the Call for Papers. An extensive and rigorous round-two paper review selected 19 papers for publication in this special issue, summarized below.

Gongmian Wang, Xing Xu, Fumin Shen, Huimin Lu, Yanli Ji, and Heng Tao Shen, "Cross-modal Dynamic Networks for Video Moment Retrieval with Text Query": In [A1], video moment retrieval is a challenging cross-modal retrieval task due to the requirement of precise temporal localization in noisy video data. For most videos, different actions within the same scene bring mutual interference noise which misleads the retrieval result. In this paper, the authors propose a Cross-modal Dynamic Networks (CDN) for video moment retrieval with text query. This fully leverages the information in the text query to reduce the noise in the visual domain with constrained inference cost. Furthermore, a new sequential frame attention mechanism is designed to extract the features of different actions within a scene. Solid experimental results show that the CDN method outperforms other State-of-the-art approaches, demonstrating its advanced status.

Date of current version April 19, 2022.
Digital Object Identifier 10.1109/TMM.2022.3159014

Tao Chen, Shui-Hua Wang, Qiong Wang, Zheng Zhang, Guo-Sen Xie, and Zhenmin Tang, "Enhanced Feature Alignment for Unsupervised Domain Adaptation of Semantic Segmentation": Due to significant domain gaps, models trained on a single labeled dataset usually fail to generalize well to another unlabeled target dataset. In [A2], the authors propose to enhance adversarial learning-based feature alignment for domain adaptive semantic segmentation. They propose a classification constrained discriminator to alleviate the training imbalance and feature distortion problems in adversarial learning. This proposed classification constrained discriminator forces the feature generator to extract domain-invariant features while maintaining the structural information useful for the semantic segmentation task. Also proposed is a hybrid framework of feature space adversarial learning with self-trained pseudo label collaboration to alleviate the classifier's source feature overfitting problem. Extensive experiments on two domain adaptation tasks demonstrate the superiority of the planned approach.

Yaoyao Zhong, Weihong Deng, Han Fang, Jiani Hu, Dongyue Zhao, Xian Li, Dongchao Wen, "Dynamic Training Data Dropout for Robust Deep Face Recognition": How to conduct effective face recognition with noise has been a valuable research subject when dealing with large-scale noisy training data. The success of large margin SoftMax loss-based models is mainly proposed for clean databases. In [A3], the authors conduct a thorough analysis of noise performance during the training process and propose a simple-yet-effective solution based on the fluctuation score of training samples, noted as dynamic training data dropout (DTDD). They leverage the information from model predictions of accumulated training epochs to distinguish the regular samples and noise. The proposed method can be combined with existing network architectures and losses. Extensive experiments on benchmark datasets empirically demonstrate that the proposed method can robustly train deep face recognition models in the presence of label noise and low-quality images.

Jingyu Hao, Guang Yang, Zhifan Gao, Jinglin Zhang, and Heye Zhang, "Annealing Genetic GAN for Imbalanced Web Data Learning": To deal with noises in partial label learning, existing approaches try to perform disambiguation either by identifying the ground-truth label or by averaging the candidate labels. However, these methods can be easily misled by the false positive noisy labels in the candidate set and fail to generalize well in testing. The authors in [A4] propose a novel paradigm called Generalized Large Margin kNN for Partial Label Learning. This paradigm aims to learn a new metric and perform disambiguation by making similarly labeled instances closer to

each other while differently labeled instances separated by a large margin. Overall, partial label learning can deal with the ambiguities caused by false-positive noisy labels in candidate label sets, further increasing the efficiency of learning from the noisy multimedia data.

Chuanyi Zhang, Qiong Wang, Guosen Xie, Qi Wu, Fumin Shen, and Zhenmin Tang, "Robust Learning from Noisy Web Images via Data Purification for Fine-Grained Recognition": It is difficult to boost fine-grained visual classification (FGVC) through enlarging manually labeled datasets due to the expert knowledge required. Previous research has primarily focused on designing complex algorithms to learn discriminative features from limited datasets. In [A5], the authors investigate a web-supervised framework for FGVC via a web dataset purification algorithm, with the aim of enlarging training sets. Specifically, authors first train a sub-center classifier to learn multiple class centers for each category and utilize the dominant sub-center for noise identification via angle distributions. Outliers in noisy samples are then detected through a rotation prediction task in a self-supervised manner. Finally, the dataset is purified by discarding outliers and relabeling other noisy images. Experiments demonstrate that the proposed method is superior to current state-of-the-art web-supervised methods.

Jing Yi and Zhenzhong Chen, "Multi-modal Variational Graph Auto-encoder for Recommendation Systems": User and item representations learned by latent factor model inherently contain uncertainty due to sparsity of user-item interactions and noise of item features. To address these challenges, [A6] presents a multi-modal variational graph auto-encoder (MVGAE) method. Specifically, the authors have designed modality-specific variational encoders that learn a Gaussian variable for each node, where the mean vector represents semantic information, and the variance vector denotes the noise level of the corresponding modality. Moreover, with the assumption of conditional independence, the modality-specific Gaussian node embeddings are fused according to the product-of-experts principle, with semantic information in each modality weighted based on the estimated uncertainty level. Extensive experiments on three public datasets – Amazon Movies, Amazon Electronics and AliShop-7C – demonstrate that MVGAE achieves competitive performance when compared with the state-of-the-art algorithms.

Zeren Sun, Huafeng Liu, Qiong Wang, Tianfei Zhou, Qi Wu, and Zhenmin Tang, "Co-LDL: A Co-training-based Label Distribution Learning Method for Tackling Label Noise": Learning with noisy labels has attracted broad attention in recent studies of computer vision due to the inferior performance caused by the memorization effect of deep neural networks. Prior works mainly adopt a low-loss sample selection strategy while neglecting the high-loss samples. In [A7], the authors propose to tackle noisy labels by integrating sample selection and label distribution learning into one unified training framework, known as Co-LDL. They simultaneously train two divergent deep networks and let them select low-loss and high-loss samples for each other. Low-loss samples are utilized conventionally, while the high-loss samples are learned in a label distribution learning manner to update network parameters and label distributions concurrently. Extensive experiments on synthetic and real-world noisy datasets demonstrate the effectiveness and superiority of the proposed Co-LDL in learning with noisy labels.

Xiuwen Gong, Dong Yuan, Wei Bao, "Generalized Large Margin kNN for Partial Label Learning": To deal with noises in partial label learning, existing approaches a to perform disambiguation either by identifying the ground-truth label or by averaging the candidate labels. However, these methods can be easily misled by the false positive noisy labels in the candidate set and fail to generalize well in testing. The authors in [A8] propose a novel paradigm called Generalized Large Margin kNN for Partial Label Learning. This paradigm aims to learn a new metric and perform disambiguation by making similarly labeled instances closer to each other while differently labeled instances are separated by a large margin. Overall, partial label learning can address the ambiguities caused by false-positive noisy labels in candidate label sets, further increasing the efficiency of learning from the noisy multimedia data.

Junya Teng, Xiankai Lu, Yongshun Gong, Xinfang Liu, Xiushan Nie, Yilong Yin, "Regularized Two Granularity Loss Function for Weakly Supervised Video Moment Retrieval": It is challenging to localize video event (video moment) from large untrimmed video data sets based on the given text descriptor (query language), particularly in weakly supervised scenarios. Many previous methods focus primarily on using video-level annotation and formulate the retrieval as a multiple instance learning issue. Reference [A9] exploits a new view for this task by properly organizing and manipulating segment-level supervision. To this end, the authors design a novel two-granularity loss function that simultaneously considers both video-level and instance-level relationships. For video level, the authors present a regularized multiple instance loss. Moreover, the authors formulate the segments-level supervision as a label correlation procedure under the noisy label and propose a temporal ensemble strategy. The proposed elegant loss function promotes the performance significantly on two well-known datasets.

Yukun Zuo, Hantao Yao, Liansheng Zhuang, Changsheng Xu, "Seek Common Ground While Reserving Differences: A Model-Agnostic Module for Noisy Domain Adaptation": Noisy domain adaptation aims to solve the problem that the source dataset contains noisy labels in domain adaptation. Previous methods handle noisy labels by selecting the small-loss samples with inconsistent predictions between two models and discarding the consistent samples, resulting in many noises contained in the selected samples. This research aims to achieve noisy domain adaptation, which indicates that the source dataset contains noisy labels in domain adaptation. Specifically, [A10] proposes an SCGWRD module by jointly considering the reliable samples with consistent predictions and inconsistent predictions. SCGWRD modules consist of both an SCG component and an RD component. The SCG component selects the reliable samples with consistent predictions for self-training, and the RD component selects the inconsistent samples to maintain the divergence between models. Overall, this is an interesting paper considering an important yet open problem.

Sijie Song, Jiaying Liu, Lilang Lin, and Zongming Guo, "Learning to Recognize Human Actions from Noisy Skeleton Data via Noise Adaptation": Most existing skeleton-based human action recognition works have been developed with relatively clean skeletons, which are not well generalizable to handle real-world noisy skeletons. Reference [A11] tries to address this new problem through 'noise adaption' instead of explicit noise modeling, negating reliance on the skeleton ground truths. According to whether the skeletons are paired, meaning the same skeleton sequence has two observations from two viewpoints, two models are proposed to reduce the impact of the noise: a regression-based model and a generation-based model. In addition, the paper also provides a new noisy skeleton dataset in which the skeletons are more similar to real-world data. The experiments conducted on this dataset and two other datasets show the proposed models consistently outperform the compared approaches.

Huafeng Liu, Haofeng Zhang, Jianfeng Lu, and Zhenmin Tang, "Exploiting Web Images for Fine-Grained Visual Recognition via Dynamic Loss Correction and Global Sample Selection": Fine-grained visual classification (FGVC) requires the annotators to have expert knowledge. Considering the cost and difficulty, creating large-scale FGVC datasets with accurate annotation is a challenging task. Therefore, training fine-grained models directly from noisy web data has attracted broad attention. In [A12], the authors propose a novel approach for removing irrelevant samples from real-world web images during training while employing useful hard examples to update the network. Concurrently, they introduce a global sampling-based model to overcome the noise rate imbalance problem common in web images. Results have shown that the proposed approach can alleviate the harmful effects of irrelevant noisy web images and hard examples to achieve better performance.

Qi Wang, Weidong Min, Qing Han, Qian Liu, Cheng Zha, Haoyu Zhao, Zitai Wei, "Inter-Domain Adaptation Label for Data Augmentation in Vehicle Re-identification": Vehicle re-identification methods often fail to achieve robust performance due to insufficient training data and domain diversities. State-of-the-art methods apply image-to-image translation or web data to achieve data augmentation. However, the construct of new datasets will not only introduce noise but also undergo a mismatch issue with the source domain. This research in [A13] aims to overcome the domain gap within a single dataset and transform the relative similarity of inter-domain subsets through a data augmentation approach, which learns domain-invariant feature representation. It proposes a multi-attribute learning network with Inter-domain Adaptation Label Smoothing Regularization (IALSR) for vehicle Re-ID. First, a multi-domain joint network (MJNet) is proposed to group several inter-domain subsets. Second, for preserving self-similarity and domain-transitivity, IALSR is designed to smooth the noise of style-transferred data.

Xiaobo Shen, Guohua Dong, Yuhui Zheng, Long Lan, Ivor W. Tsang, and Quan-Sen Sun, "Deep Co-Image-Label Hashing for Multi-label Image Retrieval": Deep supervised hashing has demonstrated the advanced learning ability of deep neural networks, and label dependencies among multi-label sets have played a crucial role in multi-label applications. In [A14], the authors propose to discover label dependency via Deep Co-Image-Label Hashing (DCILH). They treat the image and label as two views respectively and map them into a common deep Hamming space. The prototype is learned for each label and the similarities among images, labels, and prototypes are preserved during the learning process. A label-correlation aware loss on the predicted labels is further employed to exploit the label dependency. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art deep supervised hashing on large-scale multi-label image retrieval.

Zhengning Wu, Xiaobo Xia, Ruxin Wang, Jiatong Li, Tongliang Liu, "LR-SVM+: Learning Using Privileged Information with Noisy Labels": The paradigm of Learning Using Privileged Information (LUPI) always assumes that labels are annotated precisely. However, this assumption may be violated as the labels may be heavily noisy. Reference [A15] proposes a novel noise-robust SVM+ algorithm to eliminate the harmful effect of noisy labels. Specifically, the authors observe that SVM+ performs even poorer than SVM when facing noisy labels. Then they leverage privileged information to convert the optimization process into a quadratic programming problem. Furthermore, half of the small loss samples are selected to calculate the truncation threshold following the three-sigma rule, making a reliable inference. Finally, with the convex combination of the inference labels and given noisy labels, LR-SVM+ can achieve competitive performance. This paper is the first to investigate learning with privileged information and noisy labels simultaneously.

Bin Zhu, Chong-Wah Ngo, and Wing-Kwong Chan, "Learning from Web Recipe-image Pairs for Food Recognition: Problem, Baselines and Performance": Reference [A16] investigates the potential limits of using noisy web data to learn a cross-modal embedding for retrieval-based food recognition. Through a set of experiments based on known approaches, the paper finds that the model trained using noisy web recipe-image pairs manages to learn effective image features but not recipe features. The recipe features perform significantly worse than the image features trained under the single modal setting. The image features also take advantage of text-rich recipes, despite the fact that the recipes are generally noisy, demonstrating that learning image features under a cross-modal setting with recipe-image pairs is a feasible option for food recognition.

Hao-Chiang Shao, Hsin-Chieh Wang, Weng-Tai Su, and Chia-Wen Lin, "Ensemble Learning with Manifold-Based Data Splitting for Noisy Label Correction": Noises in data labels are often caused by the misplaced labels of confusing samples, which significantly degrade a model's generalization performance in supervised learning. Reference [A17] focuses on the task of noisy label correction by proposing an ensemble learning method to correct noisy labels by exploiting the local structures of feature manifolds. A nearest-neighbor-based data splitting scheme is applied to tackle noisy labels, which are locally concentrated and close to decision boundaries, by capturing the local structures of data manifolds. Hence, each sub-model can learn a coarse representation of the feature manifold, and only a few sub-models will be affected by locally-dense noisy labels. Extensive experiments and in-depth analyses on real-world datasets demonstrate the superiority of the proposed approach.

Bingwen Hu, Ping Liu, Zhedong Zheng, and Mingwu Ren, "SPG-VTON: Semantic Prediction Guidance for Multi-pose Virtual Try-on": Previous works on multi-pose virtual fitting tasks feature some problems, such as mismatch between the target clothes and the given pose, distortion of the clothes region in the try-on result, and loss of details. In [A18], the authors propose a novel end-to-end multi-pose virtual try-on framework (SPG-VTON) based on semantic prediction guidance to tackle these problems. Under the guidance of the target semantic map prediction and the clothing region mask prediction, SPG-VTON adopts a coarse-to-fine strategy to alleviate misalignment in the clothing warping process and maintain the characteristics of the desired clothes in the generated try-on image. Experimental results on two real-world data sets demonstrate that SPG-VTON is superior to current state-of-the-art methods and is robust to data noise, including background and accessory changes.

Shiji Zhou, Lianzhe Wang, Shanghang Zhang, Zhi Wang, Wenwu Zhu, "Active Gradual Domain Adaptation: Dataset and Approach": Most online web applications meet the challenge of changing environments, calling for deep learning models to adapt to the gradually changing data distribution. The phenomenon is further verified in our newly released dataset Evolving-Image-Search (EVIS), images of vehicles and digital products from 2009 to 2020 collected from a web search engine. Reference [A19] addresses this problem via active gradual domain adaptation. Specifically, the learner continually and actively selects the most informative labels and utilizes both labeled and unlabeled samples to improve the model adaptation. To this end, the authors propose the active gradual self-training (AGST) algorithm with the novel designs of active pseudo labeling and gradual semi-supervised domain adaptation. Extensive evaluations show that the proposed method can adapt well to the changing environment and yields consistently better performance than baselines.

JIAN ZHANG, *Lead Guest Editor*
University of Technology Sydney
Ultimo NSW 2007, Australia

ALAN HANJALIC, *Guest Editor*
Delft University of Technology
Delft 2628 CD, The Netherlands

RAMESH JAIN, *Guest Editor*
University of California - Irvine
CA 92697-3425 USA

XIANSHENG HUA, *Guest Editor*
DAMO Academy, Alibaba Group
Hangzhou 311121, China

SHIN'ICHI SATOH, *Guest Editor*
The National Institute of Informatics
Tokyo 101-8430, Japan

YAZHOU YAO, *Guest Editor*
Nanjing University of Science and Technology
Nanjing 210094, China

DAN ZENG, *Guest Editor*
Shanghai University
Shanghai 200444, China

APPENDIX
RELATED ARTICLES

[A1] G. Wang, X. Xu, F. Shen, H. Lu, Y. Ji, and H. Tao Shen, "Cross-modal dynamic networks for video moment retrieval with text query," *IEEE Trans. Multimedia*, vol. 24, pp. 1221–1232, 2022.

[A2] T. Chen, S.-H. Wang, Q. Wang, Z. Zhang, G.-S. Xie, and Z. Tang, "Enhanced feature alignment for unsupervised domain adaptation of semantic segmentation," *IEEE Trans. Multimedia*, vol. 24, pp. 1042–1054, 2022.

[A3] Y. Zhong, W. Deng, H. Fang, J. Hu, D. Zhao, X. Li, and D. Wen, "Dynamic training data dropout for robust deep face recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 1186–1197, 2022.

[A4] J. Hao, G. Yang, Z. Gao, J. Zhang, and H. Zhang, "Annealing genetic GAN for imbalanced web data learning," *IEEE Trans. Multimedia*, vol. 24, pp. 1164–1174, 2022.

[A5] C. Zhang, Q. Wang, G. Xie, Q. Wu, F. Shen, and Z. Tang, "Robust learning from noisy web images via data purification for fine-grained recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 1198–1209, 2022.

[A6] J. Yi, and Z. Chen, "Multi-modal variational graph auto-encoder for recommendation systems," *IEEE Trans. Multimedia*, vol. 24, pp. 1067–1079, 2022.

[A7] Z. Sun, H. Liu, Q. Wang, T. Zhou, Q. Wu, and Z. Tang, "Co-LDL: A Co-Training-based label distribution learning method for tackling label noise," *IEEE Trans. Multimedia*, vol. 24, pp. 1093–1104, 2022.

[A8] X. Gong, D. Yuan, and W. Bao, "Generalized large margin kNN for partial label learning," *IEEE Trans. Multimedia*, vol. 24, pp. 1055–1066, 2022.

[A9] J. Teng, X. Lu, Y. Gong, X. Liu, X. Nie, and Y. Yin, "Regularized two granularity loss function for weakly supervised video moment retrieval," *IEEE Trans. Multimedia*, vol. 24, pp. 1141–1151, 2022.

[A10] Y. Zuo, H. Yao, L. Zhuang, and C. Xu, "Seek common ground while reserving differences: A model-agnostic module for noisy domain adaptation," *IEEE Trans. Multimedia*, vol. 24, pp. 1020–1030, 2022.

[A11] S. Song, J. Liu, L. Lin, and Z. Guo, "Learning to recognize human actions from noisy skeleton data via noise adaptation," *IEEE Trans. Multimedia*, vol. 24, pp. 1152–1163, 2022.

[A12] H. Liu, H. Zhang, J. Lu, and Z. Tang, "Exploiting web images for fine-grained visual recognition via dynamic loss correction and global sample selection," *IEEE Trans. Multimedia*, vol. 24, pp. 1105–1115, 2022.

[A13] Q. Wang, W. Min, Q. Han, Q. Liu, C. Zha, H. Zhao, and Z. Wei, "Inter-domain adaptation label for data augmentation in vehicle re-identification," *IEEE Trans. Multimedia*, vol. 24, pp. 1031–1041, 2022.

[A14] X. Shen, G. Dong, Y. Zheng, L. Lan, I. W. Tsang, and Q.-S. Sun, "Deep co-image-label hashing for multi-label image retrieval," *IEEE Trans. Multimedia*, vol. 24, pp. 1116–1126, 2022.

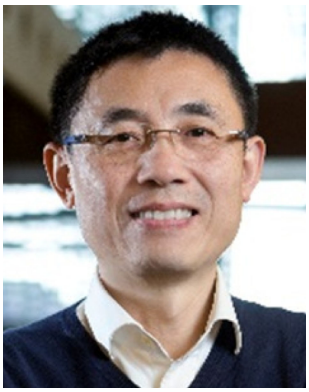
[A15] Z. Wu, X. Xia, R. Wang, J. Li, and T. Liu, "LR-SVM+: Learning using privileged information with noisy labels," *IEEE Trans. Multimedia*, vol. 24, pp. 1080–1092, 2022.

[A16] B. Zhu, C.-W. Ngo, and W.-K. Chan, "Learning from web recipe-image pairs for food recognition: Problem, baselines and performance," *IEEE Trans. Multimedia*, vol. 24, pp. 1175–1185, 2022.

[A17] H.-C. Shao, H.-C. Wang, W.-T. Su, and C.-W. Lin, "Ensemble learning with manifold-based data splitting for noisy label correction," *IEEE Trans. Multimedia*, vol. 24, pp. 1127–1140, 2022.

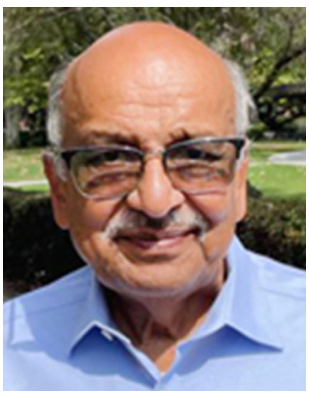
[A18] B. Hu, P. Liu, Z. Zheng, and M. Ren, "SPG-VTON: Semantic prediction guidance for multi-pose virtual try-on," *IEEE Trans. Multimedia*, vol. 24, pp. 1233–1246, 2022.

[A19] S. Zhou, L. Wang, S. Zhang, Z. Wang, and W. Zhu, "Active gradual domain adaptation: Dataset and approach," *IEEE Trans. Multimedia*, vol. 24, pp. 1210–1220, 2022.



Jian Zhang (Senior Member, IEEE) is currently a Full Professor with the Faculty of Engineering and IT and the Director of the Multimedia and Data Analytics Lab, University of Technology Sydney, Sydney, NSW, Australia. He has authored or coauthored more than 230 paper publications, book chapters and 11 approved U.S. patents. His current interests include large-scale multimedia content analysis and retrieval, 2D and 3D-based computer vision, and intelligent video surveillance systems. Dr. Zhang was the leading General Co-Chair and Technical Program Co-Chair of the International Conference on Multimedia and Expo (ICME) in 2012 and 2020, respectively, and the Technical Program Co-Chair and leading General Co-Chair of the IEEE Conference on Visual Communications and Image Processing in 2014 and 2019, respectively. He was an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY during 2006–2015 and a Member of Technical Directions Board, IEEE Signal Processing Society during January 2019–December 2020. He is also an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA.

Alan Hanjalic (Fellow, IEEE) is currently a Professor of computer science, the Head of Multimedia Computing Group, and the Head of the Intelligent Systems Department, Delft University of Technology, Delft, The Netherlands. He has coauthored more than 250 publications in his research field, which includes multimedia information retrieval and recommender systems. He was the co-recipient of the Best Paper Award at the ACM Conference on Recommender Systems (ACM RecSys) 2012, ACM International Conference on Multimedia (ACM Multimedia) 2017, and IEEE International Conference on Multimedia Big Data (IEEE BigMM) 2019. He was the Chair of the Steering Committee of the IEEE TRANSACTIONS ON MULTIMEDIA, an Associate Editor-in-Chief of the IEEE MULTIMEDIA MAGAZINE, and an Associate Editor of many scientific journals, including the IEEE TRANSACTIONS IN MULTIMEDIA, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, and *ACM Transactions on Multimedia Computing, Communications, and Applications*. He was also the General and Program Co-Chair in the organizing committees of all main conferences in the multimedia domain, including ACM Multimedia, ACM CIVR/ICMR, and IEEE ICME.



Ramesh Jain (Fellow, IEEE) is currently an Entrepreneur, Researcher, and Educator. His research interests include cybernetic systems, computer vision, artificial intelligence, data management, multimedia computing, and digital health. His current research passion is addressing health using lifestyle and environment. His dream is to radically transform health systems away from hospitals and into the hands and homes of individuals by developing a Personal Health Navigator using cybernetic principles building on the sensors, mobile, computer processing, artificial intelligence, computer vision, food computing, and medical and related technologies. He is the Founding Director of the Institute for Future Health, UCI. He is a Fellow of AAAS, ACM, AAAI, IAPR, and SPIE.

He was involved in co-founding several companies. He enjoys new challenges and likes to use technology to solve them. He is participating in addressing the biggest challenge for us all: healthy and happy life.

Xiansheng Hua (Fellow, IEEE) received the B.S. and Ph.D. degrees in applied mathematics from Peking University, Beijing, China, in 1996 and 2001, respectively. He is currently a Distinguished Engineer, the Vice President of Alibaba Group, the Head of the City Brain Lab of DAMO Academy, leading a team working on large-scale visual intelligence systems on the cloud, covering areas, such as smart city, healthcare, industrial manufacturing, agriculture, and Internet. He has authored or coauthored more than 200 research papers and has more than 60 granted patents. His research interests include big multimedia data analysis, search, and mining, as well as pattern recognition and machine learning. He is/was an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Intelligent Systems and Technology*, and *IET Smart Cities*. He was a Program Co-Chair for IEEE ICME 2013, ACM Multimedia 2012, and IEEE ICME 2012. He was one of the recipient of the 2008 MIT Technology Review TR35 Young Innovator Award for his outstanding contributions on video search. He was the recipient of the Best Paper Awards at ACM Multimedia 2007, and Best Paper Award of the IEEE Transactions on CSVT in 2014. He is also the leading General Co-Chair of ACM Multimedia 2020. Dr. Hua is an ACM Distinguished Scientist.

Shin'ichi Satoh received the B.E. degree in electronics engineering, and the M.E. and Ph.D. degrees in information engineering from the University of Tokyo, Tokyo, Japan, in 1987, 1989, and 1992, respectively. In 1992, he joined the National Center for Science Information Systems, Tokyo, Japan. Since 2004, he has been a Full Professor with the National Institute of Informatics (NII), Tokyo, Japan. From 1995 to 1997, he was a Visiting Scientist with Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. His research interests include image processing, video content analysis, and multimedia database. He is currently leading the video processing project with NII.

He was on technical program committees for several international conferences, including ACM Multimedia, ICME, ICPR, ICCV, SIGIR, and WWW. He was the General Co-Chair of International Conference on Multimedia Retrieval in 2018 (ICMR2018). He was the Program Co-Chair for Pacific-Rim Conference on Multimedia in 2004 (PCM2004), Multimedia Modeling Conference (MMM2008), International Conference on Multimedia Retrieval in 2011 (ICMR2011), ACM Multimedia 2012. He will serve as the General Co-Chair of International Conference on Multimedia Retrieval in 2022 (ICMR2022) and ACM Multimedia in 2022.

Yazhou Yao is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China, and the Ph.D. degree in computer science from the University of Technology Sydney, Sydney, NSW, Australia, in 2018, with the support of the China Scholarship Council. From July 2018 to July 2019, he was a Research Scientist with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. His research interests include multimedia processing and machine learning.

Dan Zeng received the B.S. degree in electronic science and technology and the Ph.D. degree in circuits and systems from the University of Science and Technology of China, Hefei, China. She is currently a Full Professor and the Dean of the Department of Communication Engineering and the Computer Vision and Pattern Recognition Lab, Shanghai University, Shanghai, China. Her main research interests include computer vision, multimedia analysis, and machine learning. She is an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA, an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, a TC Member of IEEE MSA, and an Associate TC Member of IEEE MMSP.