International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

LANGUAGE AND VISION

T.Srivalli , C .C Deepika , M. Gagana

I. INTRODUCTION

Natural language processing (NLP) and computer vision (CV) are two of the most essential areas in the artificial intelligence field (AI). CV is a topic of study that looks into the methods for teaching computers to perceive and interpret digital stuff like images and movies. NLP (natural language processing) is a discipline of linguistics that allows computers to process, interpret, and even synthesise human language. With the advent and development of deep learning over the last decade, there has been a continual flow of innovation and breakthroughs that convincingly push the boundaries and improve the state-of-the-art in both vision and language modelling. An noteworthy point is that research in the two areas is beginning to interact, and many previous experiences have demonstrated that this is a good thing.

In general, visual and language interactions have taken place in two dimensions: vision to language and language to vision. In the form of tags [Reference Yao, Mei, Ngo and Li1], answers [Reference Anderson2], captions [Reference Yao, Pan, Li, Qiu and Mei3–Reference Yao, Pan, Li and Mei5], and comments [Reference Li, Yao, Mei, Chao and Rui6], the former primarily recognises or describes the visual content with a set of individual words or a natural sentence. In visual content, a tag, for example, usually represents a specific object, action, or event. An answer is a statement made in response to a question concerning the facts portrayed in a photograph or video. A caption is a natural-language utterance (typically a sentence) that goes beyond tags or answers. The purpose of this programme is to convert text into an image or a video. For example, given a textual description of "this small bird has a short beak and dark stripe down the top, the wings are a mix of brown, white, and black," text-to-image synthesis' purpose is to create a bird image that fits all of the requirements.

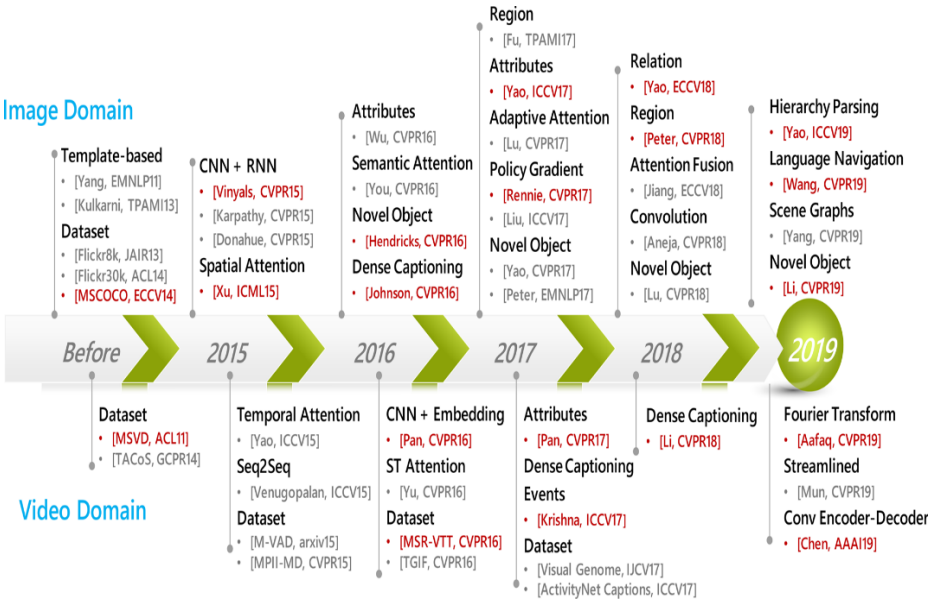
This study examines recent state-of-the-art AI advancements that improve both vision to language, particularly image/video captioning, and language to vision. Real-world deployments in both domains are also shown as good instances of how AI improves user engagement and transforms consumer experiences in industrial applications. The following is a breakdown of the remaining sections. Section II outlines the evolution of vision to language by laying out a quick road map of significant image/video captioning technologies, distilling a typical encoder–decoder structure, and summarising the results of a common benchmark. Further, the practical applications of vision to language are discussed. Section III describes the technological advances in language to vision in terms of various situations and generation methodologies, and then concludes with a summary Progress in language to image, language to video, and AI-powered applications is summarised. Finally, Section IV brings the paper to a close

II. LANGUAGE PERSPECTIVE

This section covers the evolution of vision to language (especially image/video captioning) in a variety of ways, including a road map of major approaches and benchmarks, common encoder–decoder architectures, and assessment findings of representative systems.

A) A vision-to-language road map

We've seen academics stretch the boundaries of vision to language systems (e.g. image/video captioning) over the last ten years. The road map for the techniques underpinning vision (image/video) to language, as well as the accompanying benchmarks, is shown in Figure 1. In particular, the year 2015 marks a turning point in captioning. Previously, the major stream of captioning was a template-based method in the image domain [Reference Kulkarni14,Reference Yang, Teo, Daumé III, and Aloimonos15]. The main concept is to recognise things or events in an image and then combine those words into pre-defined sentence templates such as subjective, verb, and objective. Most picture captioning datasets, such as Flickr30K and MSCOCO, will be ready to use at that time. At Deep learning-based picture captioning models were first presented in 2015. A Convolutional Neural Network (CNN) is used as an image encoder to construct image representations, and a Long Short-Term Memory (LSTM) decoder is used to generate the phrase [Reference Vinyals, Toshev, Bengio, and Erhan13]. The attention mechanism [Reference Xu16], which locates the most relevant spatial regions when anticipating each word, was also postulated at that time. Following that, the field of image captioning is rapidly expanding. Researchers devised a number of novel ideas, including augmenting image features with semantic attributes [Reference Yao, Pan, Li, Qiu, and Mei3] or visual relations [Reference Yao, Pan, Li, Qiu, and Mei4], and predicting novel objects using unpaired training data [Reference Li, Pan, Li, Qiu, and Mei4]. Reference Yao, Pan, Chao, and Mei17,Reference Yao, Pan, Li, and Mei18], and even language navigation [Reference Wang19]. Another image captioning extension is to create many sentences or phrases for a picture in order to recapitulate additional details within the image. Dense image captioning [Reference Johnson, Karpathy, and Fei-Fei20] and image paragraph generation [Reference Wang, Pan, Yao,Tang, and Mei21] are two examples of in-between methods that generate a set of descriptions or paragraphs that describe an image in more detail.

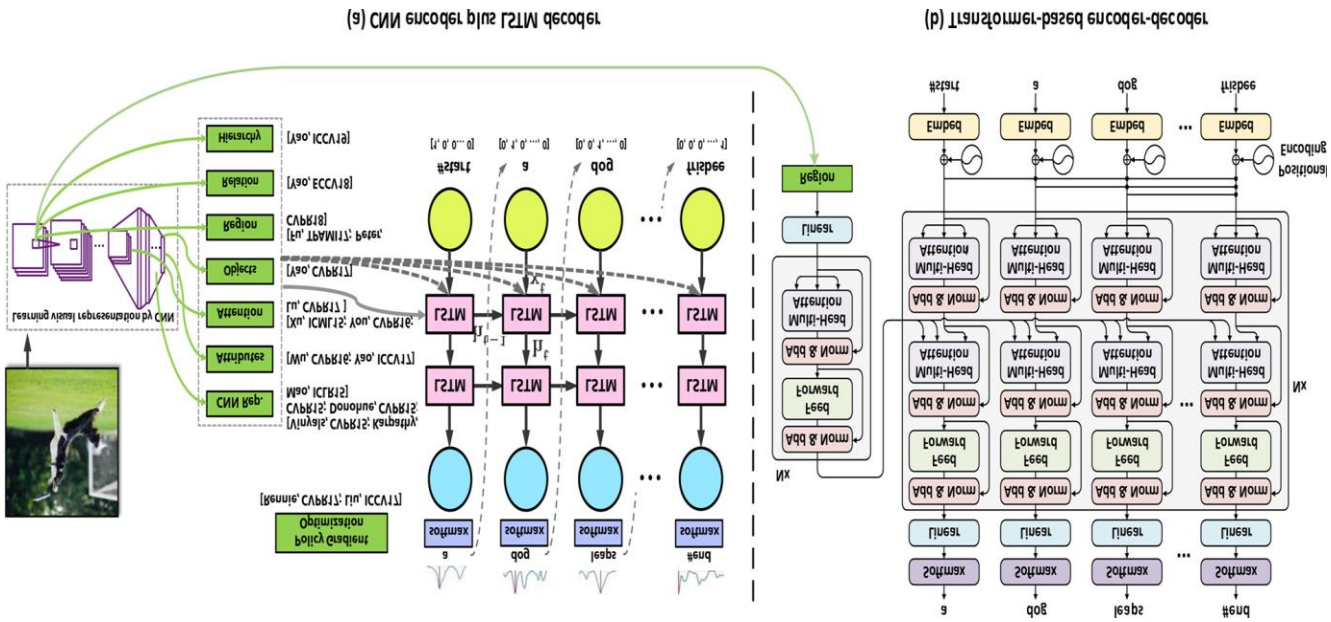


The year 2015 marks the beginning of captioning in the video domain. The researchers then begin to remould the CNN plus RNN captioning infrastructure for use in video domain captioning. To improve video captioning, a variety of strategies (such as temporal attention, embedding, and characteristics) are investigated. In

particular, Yao et al[Reference .'s Yao22] approach is one of the first to incorporate a temporal attention mechanism into a captioning framework by learning to pay attention to the most relevant frames at each decoding time step. To preserve the semantic significance between video information and the full sentence, Pan et al. [Reference Pan, Mei, Yao, Li, and Rui23] combine LSTM with semantic embedding. Pan et al. [Reference Pan, Yao, Li, and Mei24] improve captions even more. In the created text, use the model to emphasise the observed visual qualities. It's also worth noting that the MSR-VTT video captioning dataset [Reference Xu, Mei, Yao, and Rui25] was released in 2016, and it's already been downloaded by over 100 groups around the world. Recently, Aafaq et al. [Reference Aafaq, Akhtar, Liu, Gilani, and Mian26] used a quick Fourier transform to fuse all frame-level characteristics into video-level representation and improve video captioning. Another new attempt for video captioning is to use full convolutions in both encoder and decoder networks to speed up the training process [Reference Chen, Pan, Li, Yao, Chao, and Mei27]. Nonetheless, given that videos in real life are typically extensive and complex, contain several occurrences, traditional video captioning systems that generate only one caption for a video will fail to recapitulate all of the events in the film in general. As a result, dense video captioning [Reference Krishna, Hata, Ren, Fei-Fei, and Niebles28,Reference Li, Yao, Pan, Chao, and Mei29] has recently been proposed, with the eventual goal of generating a sentence for each event in the film.

B) Typical architectural structures

The bulk of modern image captioning, according to the road map of vision to language, follows the form of CNN encoder + LSTM decoder, as shown in Fig. 2. (a). Image characteristics, in particular, can be retrieved in a variety of ways given an image: (1) incorporating high-level semantic attributes into image features [Reference Yao, Pan, Li, Qiu, and Mei3]; (3) performing attention mechanism to measure the contribution of each image region [Reference Xu16]; (4) extracting region-level features [Reference Anderson2] and further exploring relation [Reference Yao, Pan, Li, and Mei4] or image hierar The image characteristics will be LSTM decoder is then used to construct the output sentence, one word at a time. During the trainings stage, the model generates the next word based on the previous ground-truth words, while during testing, the model predicts the next word using the previously generated words. Reinforcement learning [Reference Rennie, Marcheret, Mroueh, Ross and Goel8,Reference Liu, Zhu, Ye, Guadarrama and Murphy9,Reference Ren, Wang, Zhang, Lv and Li30] is commonly used to directly optimise LSTM decoders with the sentence-level reward, such as CIDEr or METEOR, in order to bridge the gap between training and testing.



Recent effort has been focused on researching Transformer-based structure [Reference Sharma, Ding, Goodman, and Soricut32] in picture captioning, inspired by the recent accomplishments of Transformer self-attention networks [Reference Vaswani31] in machine translation. The usual construction of a Transformer-based encoder–decoder is shown in Figure 2(b). Unlike CNN encoder plus LSTM decoder, which uses LSTM to describe word dependency, the Transformer-based encoder–decoder model fully exploits the attention mechanism to capture global dependencies between inputs. To model self-attention among input picture regions, N multi-head self-attention layers are placed in the encoder. A stack of N multi-head attention layers, each with a self-attention sub-layer and a cross-attention sub-layer, is present in the decoder. To be more exact, the self-attention sub-layer is used to capture word reliance The co-attention across vision (image areas from encoder) and language is also used using the sub-layer (input words).

The common paradigm in video captioning is primarily an encoder–decoder structure, similar to the norm in image captioning. 2D CNN [Reference Vinyals, Toshev, Bengio, and Erhan13] or 3D CNN [Reference Qiu, Yao, and Mei33,Reference Tran, Bourdev, Fergus, Torresani, and Paluri34] are used to encode a video into a series of frame/clip/shot features. Then, using pooling [Reference Pan, Mei, Yao, Li, and Rui23], attention [Reference Yao22], or an LSTM-based encoder [Reference Venugopalan, Rohrbach, Donahue, Mooney, Darrell, and Saenko35], all the frame-level, clip-level, or shot-level visual features are fused into video-level representations. After that, the video-level features are put into an LSTM decoder, which generates a natural sentence.

C) Application and evaluation

Evaluation. Table 1 summarises the performance of representative picture captioning algorithms on the prominent benchmark COCO [Reference Lin36] testing server. GCN-LSTM [Reference Yao, Pan, Li, and Mei4] and HIP [Reference Yao, Pan, Li, and Mei5] outperform other captioning systems on all assessment criteria, demonstrating the benefit of examining relations and hierarchical structure across image regions.

Model	Group	B@4		METEOR		ROUGE-L		CIDEr-D	
		c5	c40	c5	c40	c5	c40	c5	c40
HIP [5]	JD AI, ICCV'19	**39.3**	**71**	**28.8**	**38.1**	**59**	**74.1**	**127.9**	**130.2**
GCN-LSTM [4]	JD AI, ECCV'18	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
RFNet [7]	Tencent, ECCV'18	38	69.2	28.2	37.2	58.2	73.1	122.9	125.1
Up-Down [2]	MSR, CVPR'18	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
LSTM-A [3]	MSRA, ICCV'17	35.6	65.2	27	35.4	56.4	70.5	116	118
Watson Multimodal [8]	IBM, CVPR'17	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
G-RMI [9]	Google, ICCV'17	33.1	62.4	25.5	33.9	55.1	69.4	104.2	107.1
MetaMind/VT-GT [10]	Salesforce, CVPR'17	33.6	63.7	26.4	35.9	55	70.5	104.2	105.9
reviewnet [11]	CMU, NIPS'16	31.3	59.7	25.6	34.7	53.3	68.6	96.5	96.9
ATT [12]	Rochester, CVPR'16	31.6	59.9	25	33.5	53.5	68.2	94.3	95.8
Google [13]	Google, CVPR'15	30.9	58.7	25.4	34.6	53	68.2	94.3	94.6

Applications. There are various new applications involving vision technology and language that have recently emerged. In China, for example, captioning has been integrated into an online chatbot [Reference Pan, Qiu, Yao, Li, and Mei37,Reference Tran38] and an ai-created poem [Reference Zhou, Gao, Li, and Shum39]. Last year, JD.com began using captioning techniques for tailored product description generation, with the goal of automatically producing convincing recommendation reasons for billions of products.

III. VISION TO LANGUAGE

This section looks at "language to vision" from a different angle, i.e. visual content creation directed by linguistic inputs. In this section, we'll go over the development of the road map as well as technical developments in this field. Then we talk about the open concerns and applications, especially from the standpoint of industry.

Generating Visual Content Because "language to vision" is strongly founded in the same approaches, we quickly present the area of visual generation. We've seen significant advancements in the creation of visual material over the last few years. Visual generation has its roots in [Reference Goodfellow40], where many networks are trained in an adversarial manner together. Following works produce images in specific domains such as face [Reference Chen, Chen, Zhang, Mitchell, and Yu41–Reference Karras, Laine, and Aila43], person [Reference Ma, Jia, Sun, Schiele, Tuytelaars, and Van Gool44–Reference Song, Zhang, Liu, and Mei46], and generic domains [Reference Brock, Donahue, and Simonyan47,Reference Lui, Tschannen, In terms of inputs, the generation can alternatively be viewed as conditioning on several pieces of data, For example, noise vector [Reference Goodfellow40], semantic label [Reference Mirza and Osindero49], textual captions [Reference Reed, Akata, Yan, Logeswaran, Schiele, and Lee50], scene-graph [Reference Johnson, Gupta, and Fei-Fei51], and images [Reference Isola, Zhu, Zhou and Efros52,Reference Zhu, Park, Isola and Efro Visual generation based on natural languages is one of the most promising branches among all of these efforts, because semantics are directly incorporated into the pixel-by-pixel generation process.

A) Language to Vision Road Map

Figure 3 depicts the evolution of "language to vision" in recent years. When compared to when it was first launched in 2014, both the vision and language modalities are becoming increasingly difficult, and the outcomes are considerably more visually convincing.

The basic design is built on a conditional generative adversarial network, with encoded natural language as the conditioning input. The linguistic input is eventually translated to a visual representation with higher and better resolution after a sequence of transposed-convolutions. There are two major obstacles to overcome: (1) how to understand language input (language representation), and (2) how to match visual and textual modalities (semantic coherence between vision and language). Recent results on a single object (bottom) have been proven to be visually plausible to humans. State-of-the-art models, on the other hand, are still having trouble simulating scenarios with several objects interacting with one another.

B) Technical advancements

The success in language to vision generation is mostly based on the following technical advancements, which have become standard practices commonly accepted by the research community.

Input Preparation. The conditional version GAN was developed using the standard GAN framework [Reference Goodfellow40,Reference Mirza and Osindero49], which allows visual generation based on linguistic inputs. The conditioning data can take the shape of a tag, a sentence, a paragraph, an image, a scene-graph, or a layout. The conditioning architecture is used in almost all subsequent efforts in the "language to vision" field. Only MNIST [Reference LeCun, Bottou, Bengio, and Haffner54] digits are shown in low resolution at the time, and the conditioning information is merely a digit-label.

GAN-INT-CLS [Reference Reed, Akata, Yan, Logeswaran, Schiele, and Lee50] [Reference Reed, Akata, Yan, Logeswaran, Schiele, and Lee50] [Reference Reed, Akata, Yan, Logeswaran, Schiele, and Lee50] [Reference Reed, Akata, Yan, Logeswaran, Schiele, and Lee It bridges the gap between natural language words and image pixels for the first time. To collect visual information, the important step is to learn a text representation based on a recurrent network. The rest is generally in line with [Mirza and Osindero49]. In addition, a matching-aware discriminator is developed to ensure that the resulting image and textual input are consistent. People can make flower images by changing textual inputs, albeit the outcomes are still crude.

Architecture that is stacked. StackGAN [Reference Zhang55,Reference Zhang56] [Reference Zhang55,Reference Zhang56] [Reference Zhang55,Reference Zhang56] [Reference Zhang55,Reference Zhang56] [Reference Zhang55,Reference Zhang56] [Reference Zhang55,Reference Zhang56] [Reference

Zhang55 Unlike prior efforts, stackGAN is capable of producing realistic -pixel images by breaking the generator into many stages that are stacked sequentially. The Stage-I network only sketches the object's basic shape and colour using text representation, resulting in a low-resolution image. The Stage-II network fills in further features, like textures, based on the Stage-I output. To supplement the textual information and stabilise the training process, a conditioning augmentation strategy is also used. This stacked architecture improves the visual quality significantly over [Reference Reed, Akata, Yan, Logeswaran, Schiele, and Lee50]. Progressively-Growing GAN uses a similar concept. [Reference Aila, Karras, and Laine

Mechanism of Attention. Attention is effective in highlighting essential information, just as it is in other vision tasks. Attention is especially important in aligning keywords (language) and picture patches (vision) throughout the generating phase in "language to vision." This model is followed by two generations of attention (v1.0 and v2.0), however they differ in many specifics, such as network architecture and text encoding. AlignDraw [Reference Mansimov, Parisotto, Ba, and Salakhutdinov57] proposes painting on a canvas iteratively by looking at different words at different stages. The results, however, were not encouraging at the time. Attention 2.0, AttnGAN [Reference Xu58], and DA-GAN [Reference Ma, Fu, Chen, and Mei59], all follow a similar paradigm, but increase image quality, particularly fine-grained details, significantly.

Layout that is semantic. Recent research [Reference Song, Zhang, Liu, and Mei46, Reference Bau60, Reference Song, Zhang, Liu, and Mei46, Reference Song, Zhang, Liu, and Mei46, Reference Song, Zhang Dong, Liang, Gong, Lai, Zhu, and Yin61] have established the value of semantic layout in picture generation, where layout serves as a blueprint for the process. Semantic layout and scene-graph are introduced in language to vision to alter the language input with greater semantics. Hong et al. [Reference Hong, Yang, Choi, and Lee62] suggest first generating object bounding-boxes, then refining by estimating appearances inside each box. Johnson et al. [Reference Johnson, Gupta, and Fei-Fei51] use graph convolutions to encode object relationships from a scene graph in order to generate a layout for decoder creation. To model the spatial constraint module and contextual fusion module, Zheng et al. [Reference Zheng, Bai, Zhang, and Mei63] introduce spatial constraint module and contextual fusion module. Hinz et al. [Reference Hinz, Heinrich, and Wermter64] suggest an object pathway for multi-objects formation with complicated spatial layouts, and Hinz et al. [Reference Hinz, Heinrich, and Wermter64] propose an object pathway for multi-objects production with complex spatial layouts.

C) Achievements and applications

The following is a timeline of how "Language to Vision" evolved. On the one hand, the description of language is becoming increasingly complicated, ranging from simple words to large sentences. On the other hand, the vision aspect is growing more complicated, with the expectation of object interaction and fine-grained detail:

•Language: label sentence paragraph graph scene

•Vision: a single item surrounded by others

Image to Language Simple phrases and single-object images, such as birds [Reference Welinder65], flowers [Reference Nilsback and Zisserman66], and generic objects [Reference Russakovsky67], were the focus of early research. The visual quality has greatly improved in recent years, as seen in Fig. 3 (bottom), and some findings are convincing enough to fool human eyes.

Though a single-object image can be created well, a multi-object scene, as seen in Fig. 3, still struggles to produce realistic results (top). The introduction of semantic layout as an intermediate representation as a general trend is to reduce complexity. Machines can now generate spatially appropriate images in general, although fine-grained details are still lacking at this time.

Video to Language. Due to the large volume of information and additional temporal constraint, translating language to video is more difficult than translating language to image. There are only a few works that look into this topic. For example, Pan et al. [Reference Pan, Qiu, Yao, Li, and Mei68] use a 3D convolution process to produce video from captions. The conclusions, however, are relatively limited in terms of practical applications.

Applications- The use of "language to vision" can be divided into two categories: human-eye generation and machine-eye generation. Language to vision is already producing extremely plausible outcomes with commercial standards in several domains (e.g. face). Footnote1. People can, for example, create royalty-free facial pictures on demand. Footnote2 by manually specifying gender, hair, and eyes for games [Reference Shi, Yuan, Fan, Zou, Shi, and Liu69] or advertising. Generating data for machines and algorithms is another option. NVIDIA, for example, presented a large-scale synthetic dataset (DG-Market) for teaching people [Reference Zheng, Yang, Yu, Zheng, Yang, and Kautz70]. Models should be re-identified. Machine-generated training images are also helping some image recognition and segmentation models. However, despite the hopeful outcomes, there is still a long way to go.

IV. FINAL THOUGHTS

Human representation is based on two basic systems: vision and language. In the AI sector, combining the two into a single intelligent system has long been a goal. As we described in the study, vision to language is capable of recognising visual material and automatically providing a natural-language description, whereas language to vision is capable of characterising the internal structure in vision data and producing visual content based on language inputs. While these interactions are still in their early stages, they encourage us to learn more about the mechanisms that connect vision and language, modify real-world applications, and reconsider the integration's eventual result.

Tao Mei is a Technical Vice President at JD.com, as well as the Deputy Managing Director of JD AI Research and Director of the Computer Vision and Multimedia Lab. He was a Senior Research Manager with Microsoft Research Asia in Beijing before joining JD.com in 2018. He has over 200 publications in journals and conferences to his credit (including 12 best paper awards). He possesses over 50 patents in the United States and abroad. He is or has been a member of the Editorial Boards of IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Multimedia, ACM Transactions on Multimedia, Pattern Recognition, and other journals. He is an IEEE Fellow (2019), an IAPR Fellow (2016), and a Distinguished Scientist, ACM Fellow (2016), APSIPA Distinguished Industry Leader (2019), and IEEE Signal Processing Society Distinguished Industry Speaker (2017).

Wei Zhang is presently a Senior Researcher at Beijing-based JD AI Research. In 2015, he obtained his Ph.D. from the City University of Hong Kong's Department of Computer Science in Hong Kong, China. In 2014, he

was a visiting scholar at Columbia University's DVMM group in New York, NY, USA. He was a member of the Chinese Academy of Sciences at the time. Computer vision and visual object analysis are two of his research areas. In 2012, he was a runner-up in the TRECVID Instance Search competition, and in 2013, he got the Best Demo Award at the ACM-HK openday. He is a guest editor for TOMM, a co-chair for the ICME workshop, and a moderator for the MMM special session.

Ting Yao is a Principal Researcher of JD AI Research's Vision and Multimedia Lab in Beijing, China. His research and development efforts are focused on video comprehension, vision and language, and deep learning. He worked as a Researcher for Microsoft Research Asia in Beijing before joining JD.com. Dr. Yao participates in a number of benchmark evaluations. In international competitions such as COCO Image Captioning, Visual Domain Adaptation Challenge 2019 & 2018 & 2017, and ActivityNet Large Scale Activity Recognition Challenge 2019 & 2018 & 2017 & 2016, he is the principal designer of the top-performing multimedia analytic systems. Many accolades have been bestowed upon him, including the ACM SIGMM Outstanding Ph.D. Thesis Award in 2015, the ACM SIGMM Rising Star Award in 2019, and the IEEE TCMC Rising Star Award in 2019.