Multimedia Intelligence: When Multimedia Meets Artificial Intelligence

Wenwu Zhu, *Fellow, IEEE*, Xin Wang, *Member, IEEE*, and Wen Gao, *Fellow, IEEE*

***Abstract*—Owing to the rich emerging multimedia applications and services in the past decade, super large amount of multimedia data has been produced for the purpose of advanced research in multimedia. Furthermore, multimedia research has made great progress on image/video content analysis, multimedia search and recommendation, multimedia streaming, multimedia content delivery etc. At the same time, Artificial Intelligence (AI) has undergone a "new" wave of development since being officially regarded as an academic discipline in 1950s, which should give credits to the extreme success of deep learning. Thus, one question naturally arises: What happens when multimedia meets Artificial Intelligence? To answer this question, we introduce the concept of *Multimedia Intelligence* through investigating the mutual-influence between multimedia and Artificial Intelligence. We explore the mutual influences between multimedia and Artificial Intelligence from two aspects: i) multimedia drives Artificial Intelligence to experience a paradigm shift towards more explainability and ii) Artificial Intelligence in turn injects new ways of thinking for multimedia research. As such, these two aspects form a loop in which multimedia and Artificial Intelligence interactively enhance each other. In this paper, we discuss what and how efforts have been done in literature and share our insights on research directions that deserve further study to produce potentially profound impact on multimedia intelligence.**

***Index Terms*—Multimedia artificial intelligence, reasoning in multimedia.**

I. When Multimedia Meets AI

THE term *Multimedia* has been taking on different meanings from its first advent in 1960 s until today's common usage which refers *multimedia* to "an electronically delivered combination of media including videos, still images, audios, and texts in such a way that can be accessed interactively."[1] After evolutionary development in more than two decades [1], [2], multimedia research has also made great progress on image/video content analysis, multimedia search and recommendation, multimedia streaming, multimedia content delivery etc. The theory of Artificial Intelligence, a.k.a. AI, coming into the sight of academic researchers a little earlier in 1950s, has also experienced decades of development for various methodologies covering symbolic reasoning, Bayesian networks, evolutionary algorithms and deep learning. These two important research areas have been involving almost independently until the increasing availability of different multimedia data types enables machine learning to discover more practical models to process various kinds of real-world multimedia information and thus find its application in real-world scenarios. Therefore, a crucial question which deserves deep thinking is what will happen when multimedia and AI meet each other.

To answer this question, we propose the concept of *Multimedia Intelligence* through exploring the mutual influences between multimedia and AI. When centering multimedia around AI, multimedia drives AI to experience a paradigm shift towards more explainability, which is evidenced by the fact that a large amount of multimedia data provides great opportunities to boost the performances of AI with the help of rich and explainable information. The resulting new wave of AI can also be reflected by the plans devised by top universities or central governments for future AI. For instance, Stanford University proposed the "Artificial Intelligence 100-year (AI 100)" plan for AI in 2014 to learn how people work, live and play. Furthermore, the U.S. government later announced a proposal "Preparing for the Future of Artificial Intelligence" in 2016, setting up the "AI and Machine Learning Committee". The European Union (EU) has put forward a European approach to Artificial Intelligence, which highlights building trust in human-centric AI, including *technical robustness and safety, transparency, accountability etc.* Meanwhile, China has also established the New Generation Artificial Intelligence Development Plan emphasizing explainable and inferential AI. When centering AI around multimedia, AI in turn leads to more rational multimedia. One ultimate goal of AI is to figure out how an intelligent system can thrive in the real world and perhaps reproduce this process. The ability of perception and reasoning is one important factor that enables the survival of human in various environments. Therefore, efforts on investigation of human-like perception and reasoning in AI will lead to more inferrable multimedia with the ability to perceive and reason. However, there has been far fewer efforts focusing on this direction, i.e., utilizing the power of AI to boost multimedia through enhancing its ability of reasoning. In this

Manuscript received November 21, 2019; accepted January 9, 2020. Date of publication January 27, 2020; date of current version June 23, 2020. This work is supported by the National Natural Science Foundation of China Major Project No. U1611461. This is an invited paper in the special issue of Multimedia Computing with Interpretable Machine Learning. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yonghong Tian. *(Corresponding authors: Wen Gao; Xin Wang.)*

W. Zhu and X. Wang are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: wwzhu@tsinghua.edu.cn; xin_wang@tsinghua.edu.cn).

W. Gao is with the School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: wgao@pku.edu.cn).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.2969791

[1] [Online]. Available: https://en.wikipedia.org/wiki/Multimedia

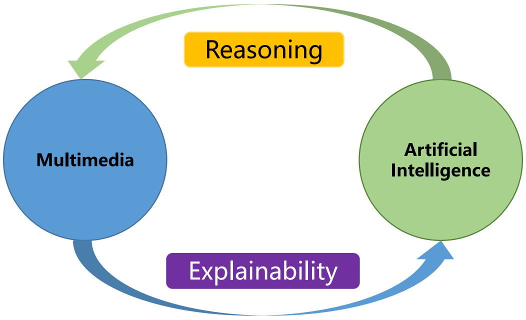


Fig. 1. The “Loop” of Multimedia Intelligence.

paper, we explore the mutual influences between multimedia and Artificial Intelligence from two aspects:

- **Center multimedia around AI:** multimedia drives AI to experience a paradigm shift towards more explainability.
- **Center AI around multimedia:** AI in turn leads to more inferrable multimedia.

Thus, *multimedia intelligence* arises with the convergence of multimedia and AI, forming the loop where multimedia and AI mutually influence and enhance each other, as is demonstrated in Fig. 1.

More concretely, given that the current AI techniques thrives with the reign of machine learning in data modeling and analysis, we discuss the bidirectional influences between multimedia and machine learning from the following two directions:

- Multimedia promotes machine learning through producing many task-specific and more explainable machine learning techniques as well as enriching the varieties of applications for machine learning.
- Machine learning boosts the inferrability of multimedia through endowing it with the ability to reason.

We summarize what have been done and analyze how well these have been done, point out what have not been done and how they possibly could be done. We further present our insights on those promising research directions that may produce profound influence on multimedia intelligence.

II. Multimedia Promotes Machine Learning

On the one hand, the multimodal essence of multimedia data drives machine learning to develop various emerging techniques such that the heterogeneous characteristics of multimedia data can be well captured and modeled [3]. On the other hand, the prevalence of multimedia data enables a wide variety of multimodal applications ranging from audio-visual speech recognition to image/video captioning and visual question answering. As is shown in Fig. 2, in this section, we discuss the ways of multimedia promoting the development of machine learning from two aspects: i) how multimedia promotes machine learning techniques and ii) how multimedia promotes machine learning applications.

A. Multimedia Promotes Machine Learning Techniques

Multimedia data contains various types of data such as image, audio and video etc., among which the single modality data has been widely studied by researchers in the past decade. However, an increasing amount of multimedia data is multimodal and heterogeneous, posing great challenges for machine learning algorithms to precisely catch the relationships among different modalities and thus appropriately deal with the multimodal data. Therefore we place our focuses on multimodal multimedia data and summarize four fundamental problems in analyzing it, i.e., *multimedia representation, multimedia alignment, multimedia fusion* and *multimedia transfer*, highlighting corresponding machine learning techniques designed to solve each of them in order to appropriately handle the various multimedia data.

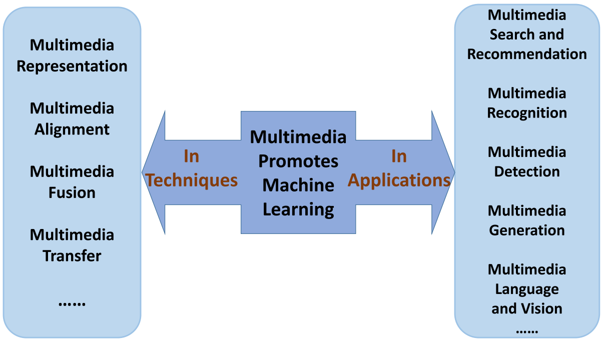


Fig. 2. Multimedia promotes machine learning.

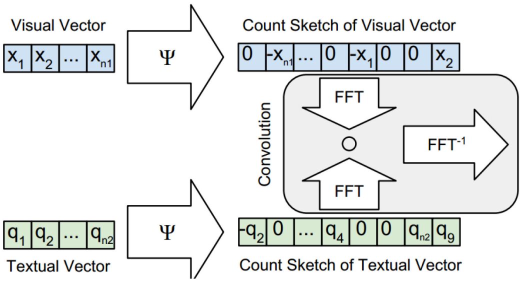

Visual Vector
x_1 x_2 ... x_{n1}
Ψ
Count Sketch of Visual Vector
0 $-x_{n1}$ ... 0 $-x_1$ 0 0 x_2
FFT
Convolution
○
FFT^{-1}
FFT
q_1 q_2 ... q_{n2}
Ψ
Textual Vector
$-q_2$ 0 ... q_4 0 0 q_{n2} q_9
Count Sketch of Textual Vector

Fig. 3. Multimodal Compact Bilinear Pooling, figure from [5].

Multimedia Representation: To represent the multimedia data, there are mainly two different categories: joint and coordinated. Joint representations combine several unimodal data into a same representation space, while coordinated representations separately process data of different modalities, but enforce certain similarity constraints on them, and make them comparable in a coordinate space. To get joint representations of multimedia data, element-wise operation, feature concatenation, fully connected layers, multimodal deep belief network [4], multimodal compact bilinear pooling [5] and multimodal convolutional neural networks [6] are leveraged or designed to combine data from different modalities. While for getting coordinated representation, a typical example is DeViSE (a Deep Visual Semantic Embedding [7]) which constructs a simple linear map from image to textual features such that corresponding annotation and image representation would have a larger inner product value between them than noncorresponding ones. Some other works also establish the coordinated space on the shared hidden layers of two unimodal auto-encoders [8], [9]. Fig. 3 shows an example of multimodal representation.

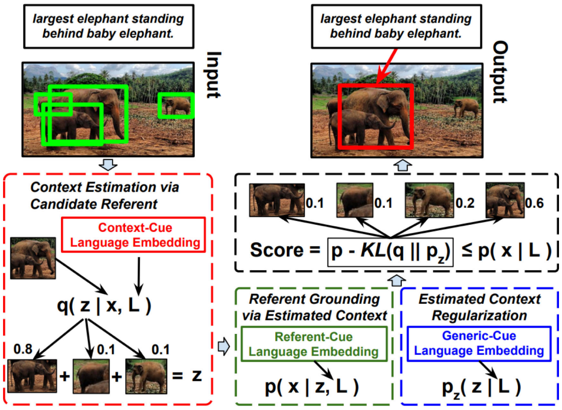

largest elephant standing behind baby elephant.
Input
largest elephant standing behind baby elephant.
Output
Context Estimation via Candidate Referent
Context-Cue Language Embedding
q(z | x, L)
0.8
0.1
0.1
= z
0.1
0.1
0.2
0.6
Score = p - KL(q || p_z) ≤ p(x | L)
Referent Grounding via Estimated Context
Referent-Cue Language Embedding
p(x | z, L)
Estimated Context Regularization
Generic-Cue Language Embedding
p_z(z | L)

Fig. 4. The variational context model for multimodal alignment, figure from [13]. Given an input referring expression and an image with region proposals, the target is to localize the referent as output. A grounding score function is developed with the variational lower-bound composed by three cue-specific multimodal modules, which is indicated by the description in the dashed color boxes.

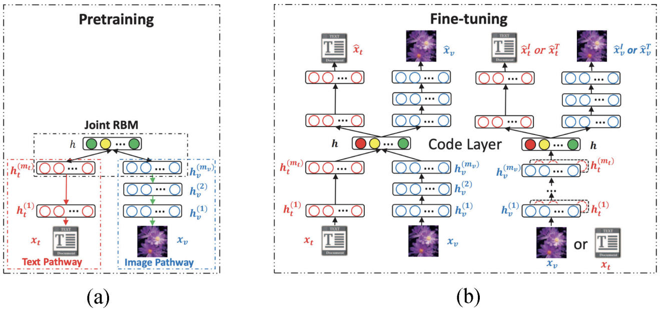


Fig. 5. (a) The multimodal DBN in pretraining and (b) The multimodal autoencoder in fine-tuning, figure from [8].

Multimedia Alignment: Multimodal multimedia data alignment is a fundamental issue for understanding multimodal data, which aims to find relationships and alignment between instances from two or more modalities. Multimodal problems such as temporal sentence localization [10]–[12] and grounding referring expressions [13], [14] are under the research field of multimodal alignment, as they need to align the sentences or phrases with the corresponding video segments or image regions. Multimodal alignment can be categorized into two main types — implicit and explicit. Baltrušaitis and Tadas *et al.* [15] categorize models whose main objective is aligning subcomponents of instances from two or more modalities as explicit multimodal alignment. In contrast, implicit alignment is used as an intermediate (normally latent) step for another task. The models with implicit alignment do not directly align data or rely on supervised alignment examples, they instead learn how to align the data in a latent manner through model training. For explicit alignment, Malmaud *et al.* [16] utilize a Hidden Markov Model (HMM) to align the recipe steps to the (automatically generated) speech transcript, Bojanowski *et al.* [17] formulate a temporal alignment problem by learning a linear mapping between visual and textual modalities, so as to automatically provide a time (frame) stamp in videos for sentences. For implicit alignment, attention mechanism [18] serves as a typical tool by telling the decoder to focus more on the targeted sub-components of the source to be translated, such as regions of an image [19], frames or segments in a video [20], [21], words of a sentence [18] and clips of an audio sequence [22] etc. Fig. 4 demonstrates an example of multimodal alignment.

Multimedia Fusion: Multimodal fusion is also one of the critical problems in multimedia artificial intelligence. It aims to integrate signals from multiple modalities together with the goal of predicting a specific outcome: a class (e.g., positive or negative) through classification, or a continuous value (e.g., population of a certain year in China) through regression. Overall, the multimodal fusion approaches can be classified into two directions [15]: model-agnostic and model-based. Model-agnostic approaches can also be split into three types: early fusion, late fusion and hybrid fusion. Early fusion integrates features from multiple modalities immediately after extraction (usually by simply concatenating their representations). Late fusion performs integration after each modality makes its own decision (e.g., classification or regression). Hybrid fusion gets consolidated outputs by combining the early fusions predictors and individual unimodal predictors together through a possibly weighted aggregation. Model-agnostic approaches can be implemented using almost any unimodal classifiers or regressors, which means the techniques they use are not designed for multimodal data. In contrast, in model-based approaches, three categories of models are designed to perform multimodal fusion: kernel-based methods, graphical models and neural networks. Multiple Kernel Learning (MKL) [23] is an extension to the kernel support vector machine (SVM) that allows different kernels to be used for data from different modalities/views. Since kernels can be seen as similarity functions between data points, the modal-specific kernel in MKL can better fuse heterogeneous data. Graphical models are another series of popular methods for multimodal fusion, which can be divided into generative methods such as coupled [24] and factorial hidden Markov models [25] alongside dynamic Bayesian networks [26] and discriminative methods such as conditional random fields (CRF) [27]. One advantage of graphical models is that they are able to exploit temporal and spatial structure of the data, making them particularly suitable for temporal modeling tasks like audio visual speech recognition. Currently, neural networks [28] have been widely used for the task of multimodal fusion. For example, long short term memory (LSTM) network [29] has demonstrated its advantages over graphical models for continuous multimodal emotion recognition [30], autoencoder has achieved satisfying performances for multimodal hashing [8], multimodal quantization [31] and video summarization [9], and convolutional neural network has been widely adopted for image-sentence retrieval tasks [6]. Although the deep neural network architectures possess the capability of learning complex patterns from a large amount of data, they suffer from the incapability of reasoning. Fig. 5 illustrates an example of multimodal fusion.

Multimedia Transfer: The problem of multimodal multimedia transfer aims at transferring useful information across different modalities with the goal of modeling a resource-poor modality by exploiting knowledge from another resource-rich

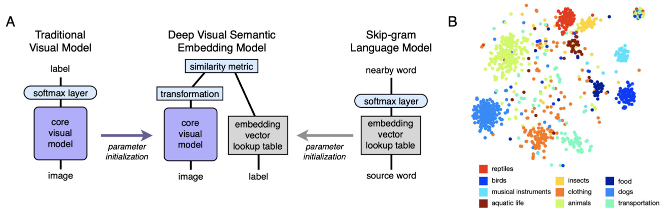


Fig. 6. (A) Left: a visual object categorization network with a softmax output layer; Right: a skip-gram language model; Center: the joint model which is initialized with parameters pre-trained at the lower layers of the other two models. (B) t-SNE visualization of a subset of the ILSVRC 2012 1 K label embeddings learned using skip-gram. Figure from [7].

modality [15]. For parallel multimodal setting which assumes modalities are from the same dataset and there is a direct correspondence between instances, transfer learning is a typical way to exploit multimodal transfer. Multimodal autoencoder [8], [28], for instance, can transfer information from one modality to another through the shared hidden layers, which not only leads to appropriate multimodal representations but also leads to better single-peak representations. Transfer learning is also feasible for non-parallel multimodal setting where modalities are assumed to come from different datasets and have overlapping categories or concepts rather than overlapping instances. This type of transfer learning is often achieved by utilizing coordinated multimodal representations. For example, DeViSE [7] uses text labels to improve image representations for classification task by coordinating CNN visual features with word2vec textual features [32] trained on separate datasets. To process non-parallel multimodal data in multimodal transfer, conceptual grounding [33] and zero shot learning [34] are two representative methodologies adopted in practice. For the hybrid multimodal setting (mixture of parallel and non-parallel data) where the instances or concepts are bridged by a third modality or a dataset, the most notable example is the Bridge Correlational Neural Network [35] which uses a pivot modality to learn coordinated multimodal representations for non-parallel data. This method can also be used for machine translation [36] and transliteration [37] to bridge different languages that do not have parallel corpora but share a common pivot language. Fig. 6 illustrates an example of multimodal transfer.

B. Multimedia Promotes Machine Learning Applications

As is discussed, the core of current AI techniques lies in the development of machine learning, therefore we will highlight several representative machine learning applications including *multimedia search and recommendation, multimedia recognition, multimedia detection, multimedia generation* and *multimedia language and vision* whose popularity should take credits from the availability of rich multimodal multimedia data.

Multimedia Search and Recommendation: Similarity search [38], [39] has always been a very fundamental research topic in multimedia information retrieval – a good similarity searching strategy requires not only accuracy but also efficiency [40]. Classical methods on similarity search are normally designed to handle the problem of searching similar contents within one single modality, e.g., searching similar texts (images) given a text (image) as query. On the other hand, the fast development of multimedia applications in recent years has created a huge number of contents such as videos, images, voices and texts which belong to various information modalities. These large volumes of multi-modal data have produced a great craving for efficient and accurate similarity search across multi-modal contents [41], [42], such as searching similar images given text queries or searching relevant texts given image queries. There have been some surveys on multi-modal retrieval and we refer interested readers to overview papers [43], [44] for more details. The fast development of Internet in the past decades has motivated the emergence of various web services with multimedia data, which drives the transformation from passive multimedia search to proactive multimedia retrieval, forming multimedia recommendation. Multimedia recommendation can cover a wide range of techniques designed for video recommendation [45], music recommendation [46], group recommendation [47] and social recommendation [48]–[50] etc. Again readers may find more detailed information about multimodal recommendation in a recent overview paper on multimodal deep analysis for multimedia [51].

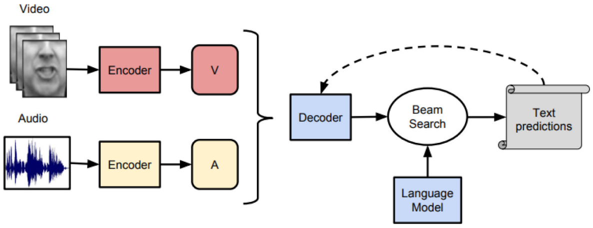


Fig. 7. Outline of the audio-visual speech recognition (AVSR) pipeline, figure from [54].

Multimedia Recognition: One of the earliest examples of multimedia research is audio-visual speech recognition (AVSR) [52]. The work was motivated by the McGurk effect [53] in which the speech perception is conducted under the visual and audio interaction of people. The McGurk effect stems from an observation that people claim to hear syllable *[da]* when seeing the film of a young talking woman where repeated utterances of syllable *[ba]* were dubbed on to lip movements for *[ga]*. These results motivate many researchers from the speech community to extend their approaches with the help of extra visual information, specifically for those from deep learning community [28], [54], [55]. Incorporating multimodal information into the speech perception procedure indeed improves the recognition performance and increase the explainability to some extend. Some others also observe that the advantages of visual information become more prominent when the audio signal ia noisy [28], [56]. The development of audio-visual speech recognition is able to facilitate a wide range of applications including speech enhancement and recognition in videos, video conferencing and hearing aids etc., especially in situations where multiple people are speaking in a noisy environment [57]. Fig. 7 presents an example of audio-visual speech recognition pipeline.

Multimedia Detection: An important research area that heavily utilizes multimedia data is human activity detection [58].

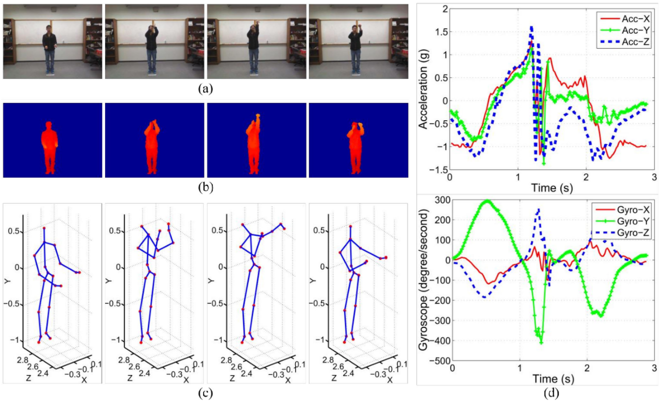


Fig. 8. An example of the multimodality data corresponding for action *basketball-shoot* : (a) color images, (b) depth images (background of each depth frame is removed), (c) skeleton joint frames, and (d) inertial sensor data (acceleration and gyro-scope signals), figure from [59].

Since human often exhibit highly complex behaviors in social activities, it is natural that machine learning algorithms resort to multimodal data for understanding and identifying human activities. Several works in deep multimodal fusion typically involve modalities such as visual, audio, depth, motion and even skeletal information [59]–[61]. Multimodal deep learning based methods have been applied to various tasks involving human activities [58], which contain action detection [62], [63] (an activity may consist of multiple shorter sequences of actions), gaze direction estimation [64], [65], gesture recognition [66], [67], emotion recognition [68], [69] and face recognition [70], [71]. The popularity of mobile smartphones with at least 10 sensors has spawned new applications involving multimodal data, including continuous biometric authentication [72], [73]. Fig. 8 demonstrates an example of multimodal detection.

Multimedia Generation: Multimodal multimedia data generation is another important aspect for multimedia artificial intelligence. Given an entity in one modality, the task is to generate the same entity in a different modality. For instance, image/video captioning and image/video generation from natural language serve as two sets of typical applications. The core ideas in multimodal generation is to translate information from one modality to another for generating contents in the new modality. Although the approaches in multimodal generation are very broad and are often modality specific, they can be categorized into two main types — example-based and generative-based [15]. Example-based methods construct a dictionary when translating between the modalities, while generative-based methods construct models that are able to produce a translation. Im2text [74] is a typical example-based method which utilizes global image representations to retrieve and transfer captions from dataset to a query image. Some other example-based methods adopt Integer Linear Programming (ILP) as an optimization framework [75], which retrieves existing human-composed phrases used to describe visually similar images, then selectively combine those phrases to generate a novel description for the query image. For generative-based methods, the encoder-decoder designs based on end-to-end trained neural networks are currently one of the most popular techniques for multimodal generation. The main idea behind such models is to first encode a source modality into a condensed vectorial representation, and then use a decoder to generate the target modality. Although encoder-decoder models are firstly used for machine translation [76], [77], they are further employed to solve image/video captioning [19], [78] and image/video/speech generation [79]–[83] problems. Fig. 9 presents an example for multimodal generation.

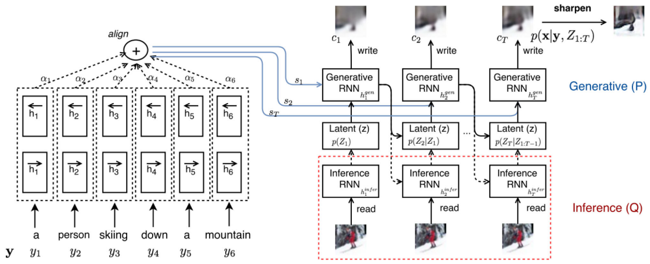


Fig. 9. AlignDRAW model for generating images by learning an alignment between the input captions and generating canvas, figure from [79]. The caption is encoded using the Bidirectional RNN (left). The generative RNN takes a latent sequence $z_{1:T}$ sampled from the prior along with the dynamic caption representation $s_{1:T}$ to generate the canvas matrix c_T, which is then used to generate the final image x (right). The inference RNN is used to compute approximate posterior Q over the latent sequence.

Multimedia Language and Vision: Another category of multimodal applications emphasize the interaction between language and vision. The most representative applications are temporal sentence localization in videos [10]–[12], image/video captioning [84]–[86] and image/video generation from natural language [79], [83], [87], [88]. Temporal sentence localization is another form of activity detection in videos, which aims to leverage natural language descriptions instead of a pre-defined list of action labels to identify specific activities in videos [10]–[12] because the complex human activities cannot be simply summarized as a constrained label set. Since natural language sentences are able to provide more detailed descriptions of the target activities, temporal boundaries can be detected more precisely with the full use of visual and textual signals [89], [90]. This can further promote a series of downstream video applications such as video highlight detection [91], video summarization [9], [92] and visual language navigation [93]. In addition, localizing natural languages in image regions is defined similarly as grounding referring expressions [13], [14]. Image/video captioning aims at generating a text description for the input image/video, which is motivated by the necesity to help visually impaired people in their daily life [94] and is also very important for content based retrieval. Therefore, the captioning techniques can be applied to many areas including biomedicine, commerce, military, education, digital libraries, and web searching [95]. Recently, some progress has also been achieved in the inverse task — image/video generation from natural language [87], [88], [96], which targets at providing more opportunities to enhance media diversity. However, both image/video captioning and generation tasks have main challenges in evaluation, i.e., how to evaluate the qualities of the predicted descriptions or generated images/videos. Fig. 10 shows an example of video captioning.

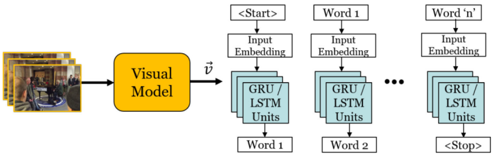


Fig. 10. A basic framework for deep learning based video captioning. A visual model encodes the video frames into a vector space. The language model takes visual vector and word embeddings as inputs to generate the sentence describing the input visual content.

III. Machine Learning Boosts Multimedia

On the one hand, exploring computer algorithm's ability for human-like perception and reasoning has always been one of top priorities in machine learning research. On the other hand, human cognition, as is illustrated in Fig. 11, can also be viewed as a cascade of perception and reasoning [97]:

- We explore our surroundings and build up our basic perceptional understanding of the world.
- We reason our perceptional understanding with our learned knowledge and obtain a deeper understanding or new knowledge.

Therefore, machine learning research focusing on studying perception and reasoning can enhance the human-like reasoning characteristics in multimedia, resulting in more inferrable multimedia.

Currently, deep learning methods can accomplish the perception parts very well: they can distinguish cats and dogs [98], identify persons [99], and answer simple questions [100]. However, they could hardly perform any reasoning: they can neither give a reasonable explanation to their perceptive prediction nor conduct explicit human-readable reasoning. Although computer algorithms are still far away from real human-like perception and reasoning, in this section we briefly review the progress of neural reasoning from the deep learning community, hoping to provide readers with a picture of what have been done in this direction.

A. Reasoning-Inspired Perception Learning

Some researchers try to equip the neural networks with reasoning ability through augmenting neural networks with reasoning-inspired layers or modules. For example, the human reasoning process may include multi-round thinking: we may repeat a certain reasoning procedure several times until reaching a certain goal. This being the case, some recurrent-reasoning layers are added to the neural network models to simulate this multi-round process. Also, relational information and external knowledge (organized as knowledge graph) are also essential for computer algorithms to gain the ability of reasoning on certain facts. These factors are also taken into account when designing deep neural networks by means of adopting Graph Neural Network [101] or Relation Network [102], [103].

Multi-Step Reasoning (RNN): The aim of multi-step reasoning is to imitate human's multi-step thinking process. Researchers insert a recurrent unit into the neural network as a multi-step reasoning module [104]–[106]. Hudson *et al.* [104] design a powerful and complex recurrent unit which is capable of meeting the definition of Recurrent Neural Network Unit and utilizing many intuitively designing such as ‘control unit,’ ‘read unit’ and ‘write unit‘ to simulate human's one-step reasoning process. Wu [105] adopt a multi-step reasoning strategy to discover step-by-step reasoning clue for visual question answering (VQA). Cadene *et al.* [106] introduce a multi-step multi-modal fusion schema to answer VQA questions. Besides, Das *et al.* [107] propose to use a multi-step retriever-reader interaction model to tackle the task of question answering. Duan *et al.* [108] uses a multi-round decoding strategy to learn better program representations of video demos. These models improve the performance significantly and claim themselves to be new state-of-the-art works for solving problems in related scenarios. However, these models are not perfect as they need more complex structures whose internal reasoning processes are even harder to interpret. Also, these methods adopt a fixed recurrent reasoning step for the sake of easy implementation, which is much less flexible than the human reasoning process.

Relational Reasoning (GNN): In addition to imitating human's multi-step reasoning process, another way of simulating human-like reasoning is utilizing graph neural network(GNN) [101] to imitate human's relational reasoning ability. Most of these works use a graph neural network to aggregate low-level perceptional features and build up enhanced features to promote the task of object detection, object tracking and visual question answering [109]–[114]. Yu *et al.* [109] and Xu *et al.* [110] use GNN to integrate features from object detection proposals for various tasks. While Narasimhan *et al.* [112] and Xiong *et al.* [113] utilize GNN as a message-passing tool to strengthen object features for visual question answering. Aside from works on image-level features, Liu *et al.* [115] and Tsai *et al.* [116] build graphs on spatial-temporal data for video social relations detection. Duan *et al.* [117] use relational data to improve 3D point cloud classification performances as well as increase the model's interpretability. In their work, an object can be seen as a combination of several sub-objects who together with their relations define the object. For example, a bird can be seen as a complex integration of sub-objects such as ‘wings,’ ‘legs,’ ‘head,’ ‘body’ and their relations, which is believed to be capable of improving the model performance and interpretability. Besides, Wen *et al.* [118] take relations among multiple agents into consideration for the task of multi-agent reinforcement learning, and Chen *et al.* [119] propose a two-stream network that combines convolution-based and graph-based model together.

Attention Map and Visualization: A lot of works use the attention maps as a way of reasoning visualization or interpretation. These attention maps, to some extent, validate the reasoning ability of the corresponding methods. In particular, Mascharka *et al.* [120] propose to use attention map as a visualization and reasoning clue. Cao *et al.* [121] use the dependency tree to guide the attention map for VQA task. Fan *et al.* [122] resort to latent attention map to improve multi-model reasoning tasks.

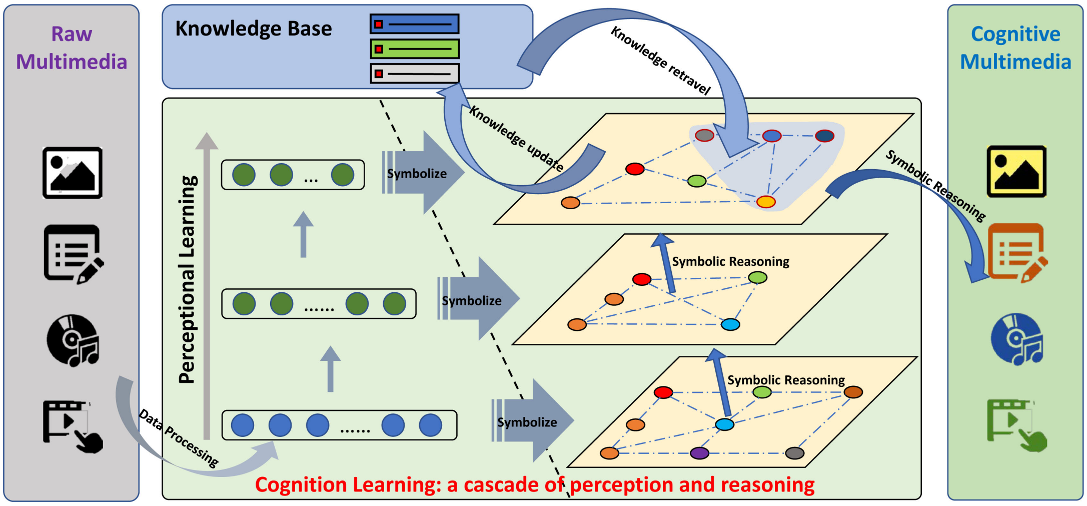


Fig. 11. Human-like Cognition.

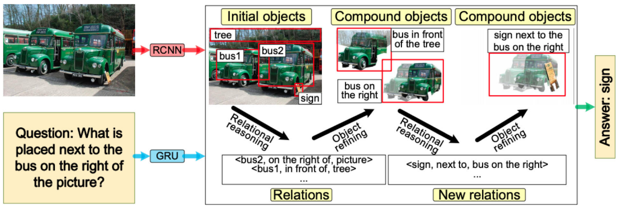


Fig. 12. A multi-step reasoning model pipeline for visual question answering [105].

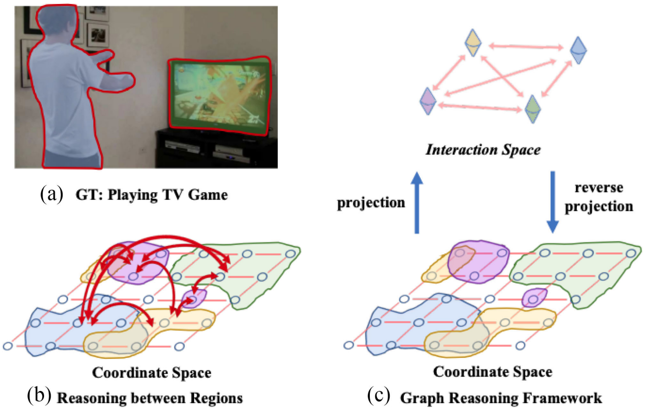


Fig. 13. Use GNN as a reasoning tool for the task of object detection [114].

B. Perception-Reasoning Cascade Learning

On the one hand, quite a few efforts have been devoted to integrating the ability to reason into deep neural networks (DNN). On the other hand, others try to decouple DNN's powerful low-level representation ability and cascade the process of perception to simulate high-level human-readable cognition, aiming at true AI [97].

Neural Modular Network: Neural module network (NMN) is first proposed by Andreas *et al.* [123] and further finds its applications in visual reasoning tasks. The main idea of NMNs is to dynamically assemble instance-specific computational graphs with a collection of pre-defined neural modules, thus enabling personalized heterogeneous computations for each input instance. The neural modules are designed with specific functions, e.g., `Find`, `Relate`, `Answer` etc., and typically assembled into a hierarchical tree structure on the fly according to different input instances.

The motivation of NMN comes from two observations:

1) Visual reasoning is inherently compositional.
2) Deep neural networks have powerful representation capacities.

The compositional property of NMN allows us to decompose visual reasoning procedure into several shareable, reusable primitive functional modules. Afterwards, deep neural networks can be used to implement these primitive functional modules as neural modules effectively. The merits of modeling visual capability as hierarchical primitives are manifold. First, it is possible to distinguish low-level visual perception from higher-level visual reasoning. Second, it is able to maintain the compositional property of the visual world. Third, the resulting models are more interpretable compared with holistic methods, potentially benefiting the development of human-in-the-loop multimedia intelligence in the future.

Visual question answering (VQA) task is a great testbed for developing computer algorithms' visual reasoning abilities. The most widely-used VQA datasets [100], [124] emphasize much more on visual perception rather than visual reasoning, motivating the existences of several challenging datasets for multi-step, compositional visual reasoning [125], [126]. The CLEVR dataset [125] consists of a set of compositional questions over synthetic images rendered with only 3 classes of objects and 12 different properties (e.g., large blue sphere), while the GQA dataset [126] operates over real images with a much larger semantic space and more diverse visual concepts.

As the earliest work, Andreas *et al.* [123] propose the NMNs to compose heterogeneous, jointly-trained neural *modules* into deep networks. They utilize dependency parsers and handwritten rules to generate module *layout*, according to which they then assemble a deep network using a small set of modules to

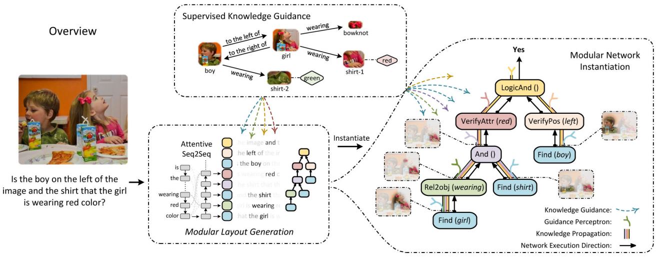
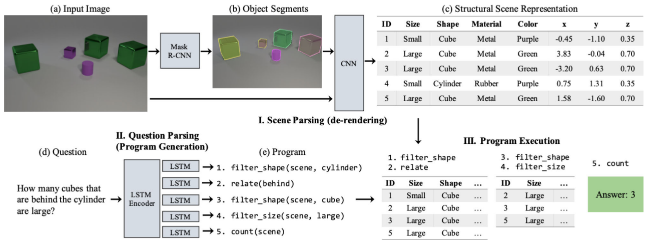


Fig. 14. An overview of the Perceptual Visual Reasoning (PVR) model, figure from [131].


(a) Input Image
Mask R-CNN
(b) Object Segments
CNN
(c) Structural Scene Representation
I. Scene Parsing (de-rendering)
II. Question Parsing (Program Generation)
(d) Question
How many cubes that are behind the cylinder are large?
LSTM Encoder
LSTM
LSTM
LSTM
LSTM
LSTM
(e) Program
III. Program Execution
Answer: 3

ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

```
1. filter_shape(scene, cylinder)
2. relate(behind)
3. filter_shape(scene, cube)
4. filter_size(scene, large)
5. count(scene)
```

```
1. filter_shape
2. relate
```

ID	Size	Shape	...
1	Small	Cube	...
2	Large	Cube	...
3	Large	Cube	...
5	Large	Cube	...

```
3. filter_shape
4. filter_size
```

ID	Size	...
2	Large	...
3	Large	...
5	Large	...

```
5. count
```

Fig. 15. Neural symbolic reasoning for visual question answering, figure from [132]. The image and question are first symbolized using neural networks, and then the symbolized representations are passed into a reasoning tool to obtain the answer.

answer visual questions. Later work on dynamic module networks (D-NMNs) [127] learns to select the optimal layout from a set of candidate layouts which are automatically generated using hand-written rules. Instead of relying on off-the-shelf parsers to generate layouts, Hu *et al.* [128] and Johnson *et al.* [129] concurrently propose to formulate the layout prediction problem as a sequence-to-sequence learning problem. Both models can predict network layouts while simultaneously learn network parameters end-to-end using a combination of REINFORCE and gradient descent. Notably, the model proposed by Johnson [129] designs fine-grained highly-specialized modules for CLEVR dataset [125], e.g., `filter_rubber_material`, which hard-code textual parameters in module instantiation. In contrast, the End-to-End Module Networks (N2NMNs) model proposed by Hu *et al.* [128] designs a set of general modules, e.g., `Find`, `Relocate`, that accept soft attention word embeddings as textual parameters. In the later work by Hu *et al.* [130] – Stack Neural Module Network (Stack-NMN), instead of making discrete choices on module layouts, the authors make the layout soft and continuous with a fully differentiable stack structure. Mascharka *et al.* [120] proposes a Transparency by Design network (TbD-net), which uses fine-grained modules similar to [129] but redesigns each module according to the intended function. This model not only demonstrates near-perfect performance on CLEVR dataset [125] but also shows visual attention that provides interpretable insights into model behavior.

Although these modular networks demonstrate near-perfect accuracy and interpretability on synthetic images, it remains challenging to perform comprehensive visual reasoning on real-world images. Recently, Li *et al.* [131] propose the Perceptual Visual Reasoning (PVR) model for compositional and explainable visual reasoning on real images, as shown in Figure 14. The authors design a rich library of universal modules ranging from low-level visual perception to high-level logic inference. Meanwhile, each module in the PVR model is capable of perceiving external supervision from guidance knowledge, which helps the modules to learn specialized and decoupled functionalities. Their experiments on the GQA dataset demonstrate that the PVR model can produce transparent, explainable intermediate results in the reasoning process.

Neural-Symbolic Reasoning: In addition to organizing modular neural networks with linguistic layout, neural-symbolic reasoning is also an advanced and promising direction which is motivated by the cognitive models from cognitive science, artificial intelligence, and psychology as well as the development of cognitive computational systems integrating machine learning and automated reasoning. Garcez *et al.* [133] introduce the basic idea of neural-symbolic reasoning: Neural Networks are first used to learn low-level perceptional understanding of the scene, and then the learned results are regarded as discrete symbols to conduct reasoning under any reasoning techniques. Most recently, Yi *et al.* [132] explore the ability of neural-symbolic reasoning under visual question answering. The task of visual question answering is disentangled into visual concept detection, language to program translation, and program execution. By learning visual symbolic representations and language symbolic representations, neural-symbolic reasoning is able to answer the visual question by 'executing' the learned language symbolic codes on the visual symbolic graph under a pre-designed program executor.

Neural-symbolic reasoning attracts lots of research interests recently for its capability of utilizing DNN‘s powerful feature representation ability and simulating human's high level reasoning and cognition. However, the ad-hoc program designer and complex program executor used by neural-symbolic reasoning severely restrict its performances and developing better program designers and executors really deserves more investigations in the future.

IV. Future Research and Directions

A. *Multimedia Turing Test*

In this paper, we introduce the concept of multimedia intelligence and present a loop (as is illustrated in Fig. 1) between multimedia and AI in which they interactively co-influence each other. As we mentioned before, the half loop from multimedia to AI (machine learning) has been well studied by recent research while the other half of the loop from AI (machine learning) to multimedia has been far less investigated, which indicates the incompleteness in the loop. We consider *multimedia Turing test* as a promising way towards completing the the loop. Multimedia Turing test consists of visual Turing test (visual and text), audio Turing test (audio and text) etc., where the Turing test is conducted on multiple multimedia modalities. We take visual Turing test as an example in this section and argue that it will be similar for other members in multimedia Turing test. Passing visual Turing test which aims to evaluate the computer algorithm's ability of human-level concept learning may serve as a further step to enhance the human-like reasoning for multimedia. The introduction of visual Turing test is originally motivated by the ability of humans to understand an image and even tell a story about it. In a visual Turing test, both the test machine and human are given an image and a sequence of questions that follow a natural story line which similar to what humans do when they look at a picture. If we human fail to distinguish between the person and machine in the test by checking their answers to the sequence of questions given an image, then it is fair to conclude that the machine passes the visual Turing test. It is obvious that passing a visual Turing test requires human-like reasoning ability.

B. Explainable Reasoning in Multimedia

For future work, exploring more explainable reasoning procedures for multimedia will be one important research direction deserving further investigations. One simple way is to enrich deep neural networks with reasoning-characteristics by utilizing other reasoning characteristics to augment deep neural networks. We should equip deep neural networks with more and better reasoning-augmented layers or modules, these modules would improve DNN's representation ability. For example, various multimedia objects can be connected by heterogeneous networks and thus be modeled through GNNs. Then it will be promising to combine the ability of relational reasoning in GNN with human-like multi-step reasoning to develop a new GNN framework with more powerful reasoning ability. Taking a deeper thinking, the most attractive part of human-like cognition learning (perception-reasoning cascade learning in Fig. 11) is that the reasoning process is transparent and explainable, which means we know how and why our models would act toward a certain scenario. Thus designing more powerful reasoning models with the help of first-order logic, logic programming language, or even domain-specific language and more flexible reasoning technique deserves further investigation. Also, the automation of program language designing and program executor can enable the adoption of neural-symbolic reasoning in more complex scenarios, which is another promising way towards explainable reasoning in multimedia. Last, given that current neural networks and the reasoning modules are optimized separately, the incorporation of neural network and reasoning through a joint-optimizing framework plays an important role in achieving the goal of explainable reasoning in multimedia.

C. AutoML and Meta-Learning

Automated Machine Learning (AutoML) and Meta-learning are exciting and fast-growing research directions to the research community in both academia and industry. AutoML targets at automating the process of the applying end-to-end machine learning models to real-world problems. The fundamental idea of AutoML is enabling a computer algorithm to automatically adapt to different data, tasks and environments, which is exactly what we human are good at. Although some efforts have been made on developing AutoML models through exploring Neural Architecture Search (NAS) for deep neural networks and Hyper-Parameter Optimization (HPO) for general machine learning models, they are still far from achieving a level comparable with human, let alone applying the core idea of AutoML to multimedia data which are multimodal in essence.

Meta-learning, i.e., learning to learn, aims at extracting and learn a form of general knowledge from different tasks that can be used by various other tasks in the future, which is also a unique characteristic possessed by human. Existing literature on meta-learning mainly focus on measuring the similarities across different data or tasks and attempting to remember (keep) previous knowledge as much as possible with the help of extra storage. It is still a long way to go for the current algorithms to summarize and further sublime previous data/knowledge into a more general form of knowledge shared across various tasks in a human-like manner.

Therefore, applying the ideas of AutoML and meta-learning on multimodal multimedia problems and developing the ability of human-like task/environment-adaptation and general knowledge sublimation is another key ingredient for advancing the new wave of AI.

D. Digital Retinas

Last but not least, as we point out in Fig. 11, there are actually no strict boundary between perception and reasoning during the process of human cognition — it is possible that we perceive and reason at the same time. Therefore, developing some prototype systems simulating this process may push the loop of multimedia intelligence one giant step towards a perfect closure.

Take the real-world video surveillance systems as an example, video streams in the current systems are firstly captured and compressed at the cameras, and then transmitted to the backend severs or cloud for big data analysis and retrieval. However, it is recognized that compression will inevitably affect visual feature extraction, consequently degrading the subsequent analysis and retrieval performance. More importantly, it is impractical to aggregate all video streams from hundreds of thousands of cameras for big data analysis and retrieval. The idea of human-like cognition learning, i.e., cascade of perception and reasoning, can be adopted as one possible solution. Let us image that we design a new framework of camera, which is called digital retina. This new digital retina is inspired by the fact that a biologic retina actually encodes both pixels and features, while the downstream areas in the brain receive not a generic pixel representation of the image, but a highly processed set of extracted features. Under the digital retina framework, a camera is typically equipped with a globally unified timer and an accurate positioner, and can output two streams simultaneously, including a compressed video stream for online/offline viewing and data storage, and a compact feature stream extracted from the original image/video signals for pattern recognition, visual analysis and search. There are three key technologies to enable the digital retina, including analysis-friendly scene video coding, visual feature compact descriptor, and joint compression of both visual content and features. By real-time feeding only the feature streams into the cloud center, these cameras thus are able to form a large-scale brain-like vision system for the smart city. There will be no doubt that successfully possessing such a brain-like system can dramatically move the current multimedia research towards a more rational and human-like manner.

V. Conclusion

In this paper, we reveal the convergence of multimedia and AI in the "big data" era. We present the novel concept of *Multimedia Intelligence* which explores the co-influence between multimedia and AI. The exploration includes the following two directions:

1) Multimedia drives AI towards more explainability.
2) AI in turn boosts multimedia to be more inferrable.

These two directions form a loop of multimedia intelligence where multimedia and AI enhance each other in an interactive and iterative way. We carefully study the circles in the loop, in particular, investigating how multimedia promotes machine learning and how machine learning in turn boosts multimedia. Last but not least, we summary what have been done in the loop already and point out what needs to be done to complete the loop, followed by our thought on several future research directions deserving further study for multimedia intelligence.

Acknowledgment

The Authors would like to thank Xuguang Duan, Guohao Li and Yitian Yuan for providing relevant materials and valuable opinions.

References

[1] B. Li, Z. Wang, J. Liu, and W. Zhu, "Two decades of internet video streaming: A retrospective view," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 9, 2013, Art. no. 33.

[2] L. Zhang and Y. Rui, "Image search from thousands to billions in 20 years," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 9, 2013, Art. no. 36.

[3] M. Cord and P. Cunningham, *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*. Berlin, Germany: Springer, 2008.

[4] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," in *Proc. Int. Conf. Mach. Learn. Workshop*, vol. 79, 2012.

[5] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Nov. 2016, pp. 457–468.

[6] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 2623–2631.

[7] A. Frome *et al.*, "DeViSE: A deep visual-semantic embedding model," in *Advances Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[8] D. Wang, P. Cui, M. Ou, and W. Zhu, "Deep multimodal hashing with orthogonal regularization," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 2291–2297.

[9] Y. Yuan, T. Mei, P. Cui, and W. Zhu, "Video summarization by learning deep side semantic embedding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 1, pp. 226–237, Jan. 2017.

[10] L. Anne Hendricks *et al.*, "Localizing moments in video with natural language," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5803–5812.

[11] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "TALL: Temporal activity localization via language query," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5267–5275.

[12] Y. Yuan, T. Mei, and W. Zhu, "To find where you talk: Temporal sentence localization in video with attention based location regression," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, pp. 9159–9166.

[13] H. Zhang, Y. Niu, and S.-F. Chang, "Grounding referring expressions in images by variational context," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 4158–4166.

[14] D. Liu, H. Zhang, Z. J. Zha, and F. Wang, "Referring expression grounding by marginalizing scene graph likelihood," 2019, *arXiv:1906.03561*.

[15] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[16] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabinovich, and K. Murphy, "What's cookin'? Interpreting cooking videos using text, speech and vision," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2015, pp. 143–152.

[17] P. Bojanowski *et al.*, "Weakly-supervised alignment of video with text," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 4462–4470.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Representations*, Jan. 2015, pp. 1–13.

[19] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, 2015, pp. 3156–3164.

[20] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 4507–4515.

[21] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4584–4593.

[22] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 4960–4964.

[23] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *J. Mach. Learn. Res.*, vol. 12, no. Jul, pp. 2211–2268, 2011.

[24] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech Recognit," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. 2, pp. II–2013.

[25] Z. Ghahramani and M. I. Jordan, "Factorial hidden markov models," in *Proc. Advances Neural Inf. Process. Syst.*, 1996, pp. 472–478.

[26] A. Garg, V. Pavlovic, and J. M. Rehg, "Boosted learning in dynamic bayesian networks for multimodal speaker detection," *Proc. IEEE*, vol. 91, no. 9, pp. 1355–1369, 2003.

[27] J. D. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 282–289.

[28] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image Vision Comput.*, vol. 31, no. 2, pp. 153–163, 2013.

[31] X. Wang, W. Zhu, and C. Liu, "Semi-supervised deep quantization for cross-modal search," in *Proc. 27th ACM Int. Conf. Multimedia*. ACM, 2019, pp. 1730–1739.

[32] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Advances Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[33] M. Baroni, "Grounding distributional semantics in the visual world," *Lang. Linguistics Compass*, vol. 10, no. 1, pp. 3–13, 2016.

[34] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. Advances Neural Inf. Process. Syst.*, 2013, pp. 935–943.

[35] J. Rajendran, M. M. Khapra, S. Chandar, and B. Ravindran, "Bridge correlational neural networks for multilingual multimodal representation learning," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, Jun. 2016, pp. 171–181.

[36] P. Nakov and H. T. Ng, "Improving statistical machine translation for a resource-poor language using related resource-rich languages," *J. Artif. Intell. Res.*, vol. 44, pp. 179–222, 2012.

[37] M. M. Khapra, A. Kumaran, and P. Bhattacharyya, "Everybody loves a rich cousin: An empirical study of transliteration through bridge languages," in *Proc. Human Lang. Technol.: Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 420–428.

[38] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient similarity search in sequence databases," in *Proc. Int. Conf. Found. Data Org. Algorithms*, 1993, pp. 69–84.

[39] P. Ciaccia, M. Patella, and P. Zezula, "M-tree: An efficient access method for similarity search in metric spaces," in *Proc. 23 rd VLDB Conf.*, Athens, Greece. Citeseer, 1997, pp. 426–435.

[40] A. Gionis, P. Indyk, R. Motwani *et al.*, "Similarity search in high dimensions via hashing," in *Proc. 25th Int. Conf. Very Large Data Bases*, vol. 99, no. 6, 1999, pp. 518–529.

[41] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos, "Bridging the Gap: Query by semantic example," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 923–938, Aug. 2007.

[42] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*. ACM, 2010, pp. 251–260.

[43] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," 2016, *arXiv:1607.06215*.

[44] J. Wang, T. Zhang, J. song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 769–790, May 2017.

[45] S. Yu, X. Wang, W. Zhu, P. Cui, and J. Wang, "Disparity-preserved deep cross-platform association for cross-platform video recommendation," in *Proc. 28th Int. Joint Conf. Artif. Intell., IJCAI-19*, 7 2019, pp. 4635–4641. [Online]. Available: https://doi.org/10.24963/ijcai.2019/644

[46] A. Van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Proc. Advances Neural Inf. Process. Syst.*, 2013, pp. 2643–2651.

[47] X. Wang *et al.*, "Recommending groups to users using user-group engagement and time-dependent matrix factorization," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 1331–1337.

[48] X. Wang, W. Lu, M. Ester, C. Wang, and C. Chen, "Social recommendation with strong and weak ties," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manag.*, 2016, pp. 5–14.

[49] X. Wang, S. C. Hoi, M. Ester, J. Bu, and C. Chen, "Learning personalized preference of strong and weak ties for social recommendation," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 1601–1610.

[50] X. Wang, W. Zhu, and C. Liu, "Social recommendation with optimal limited attention," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2019, pp. 1518–1527.

[51] W. Zhu, X. Wang, and H. Li, "Multi-modal deep analysis for multimedia," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2019.2940647.

[52] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Commun. Mag.*, vol. 27, no. 11, pp. 65–71, Nov. 1989.

[53] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, 1976, Art. no. 746.

[54] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018, to be published.

[55] S. Petridis *et al.*, "End-to-end audiovisual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6548–6552.

[56] M. Gurban, J.-P. Thiran, T. Drugman, and T. Dutoit, "Dynamic modality weighting for multi-stream HMM inaudio-visual speech recognition," in *Proc. 10th Int. Conf. Multimodal Interfaces*. ACM, 2008, pp. 237–240.

[57] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graphics*, vol. 37, no. 4, pp. 112:1–112:11, 2018.

[58] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[59] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 168–172.

[60] S. Escalera *et al.*, "ChaLearn looking at people challenge 2014: Dataset and results," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 459–473.

[61] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive multimodal human action database," in *Proc. IEEE Workshop Appl. Comput. Vision*, 2013, pp. 53–60.

[62] P. Natarajan *et al.*, "Multimodal feature fusion for robust event detection in web videos," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 1298–1305.

[63] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 1961–1970.

[64] S. S. Mukherjee and N. M. Robertson, "Deep head pose: Gaze-direction estimation in multimodal video," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2094–2107, Nov. 2015.

[65] D. Lian *et al.*, "Multiview multitask gaze estimation with deep convolutional neural networks," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30, no. 10, Sep. 2018, pp. 3010–3023.

[66] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "ModDrop: Adaptive multi-modal gesture recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1692–1706, Aug. 2016.

[67] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Trans. Pattern Anal.Mach. Intell.*, vol. 38, no. 8, pp. 1583–1597, Aug. 2016.

[68] S. E. Kahou *et al.*, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[69] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining*, 2016, pp. 439–448.

[70] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2049–2058, Nov. 2015.

[71] W. Zhang, Y. Zhang, L. Ma, J. Guan, and S. Gong, "Multimodal learning for facial expression recognition," *Pattern Recognit.*, vol. 48, no. 10, pp. 3191–3202, 2015.

[72] Z. Sitová *et al.*, "HMOG: New behavioral biometric features for continuous authentication of smartphone users," *IEEE Trans. Inf. Forensics Secur.*, vol. 11, no. 5, pp. 877–892, May 2015.

[73] P. T. Schultz and R. A. Sartini, "Method and system for multi-factor biometric authentication," Apr. 2016, U.S. Patent 9,323,912.

[74] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2Text: Describing images using 1 million captioned photographs," in *Advances Neural Inf. Process. Syst.*, 2011, pp. 1143–1151.

[75] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, "Collective generation of natural image descriptions," in *Proc. 50th Annu. Meet. Assoc. Comput. Linguistics: Long Papers-Vol. 1*, 2012, pp. 359–368.

[76] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[77] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1700–1709.

[78] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol*, 2015, pp. 1494–1504.

[79] E. Mansimov, E. Parisotto, J. L. Ba and R. Salakhutdinov, "Generating images from captions with attention," 2015, *arXiv:1511.02793*.

[80] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 1060–1069.

[81] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2405–2413.

[82] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[83] Y. Liu, X. Wang, Y. Yuan, and W. Zhu, "Cross-modal dual learning for sentence-to-video generation," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1239–1247.

[84] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4651–4659.

[85] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6504–6512.

[86] X. Duan *et al.*, "Weakly supervised dense event captioning in videos," in *Proc. Advances Neural Inf. Process. Syst.*, 2018, pp. 3059–3069.

[87] T. Qiao, J. Zhang, D. Xu, and D. Tao, "MirrorGAN: Learning text-to-image generation by redescription," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 1505–1514.

[88] Y. Pan, Z. Qiu, T. Yao, H. Li, and T. Mei, "To create what you tell: Generating videos from captions," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 1789–1798.

[89] J. Chen, X. Chen, L. Ma, Z. Jie, and T.-S. Chua, "Temporally grounding natural sentence in video," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 162–171.

[90] D. Zhang, X. Dai, X. Wang, Y.-F. Wang, and L. S. Davis, "Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 1247–1257.

[91] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 982–990.

[92] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing web videos using titles," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 5179–5187.

[93] P. Anderson *et al.*, “Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 3674–3683.

[94] J. P. Bigham *et al.*, “VizWiz: Nearly real-time answers to visual questions,” in *Proc. 23nd Annual ACM Symp. User Interface Softw. Technol*, 2010, pp. 333–342.

[95] M. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, “A comprehensive survey of deep learning for image captioning,” *ACM Comput. Surveys*, vol. 51, no. 6, 2019, Art. no. 118.

[96] H. Zhang *et al.*, “StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5907–5915.

[97] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, and L. De Raedt, “DeepProbLog: Neural probabilistic logic programming,” in *Advances Neural Inf. Process. Syst.*, 2018, pp. 3749–3759.

[98] J. Deng *et al.*, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.

[99] L. Zhao, X. Li, Y. Zhuang, and J. Wang, “Deeply-learned part-aligned representations for person re-identification,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 3219–3228.

[100] S. Antol *et al.*, “VQA: Visual question answering,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 2425–2433.

[101] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, “The graph neural network model,” *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Dec. 2008.

[102] A. Santoro *et al.*, “A simple neural network module for relational reasoning,” in *Proc. Advances Neural Inf. Process. Syst.*, 2017, pp. 4967–4976.

[103] R. Palm, U. Paquet, and O. Winther, “Recurrent relational networks,” in *Proc. Advances Neural Inf. Process. Syst.*, 2018, pp. 3368–3378.

[104] D. A. Hudson and C. D. Manning, “Compositional attention networks for machine reasoning,” in *Proc. 6th Int. Conf. Learn. Representations*, 2018, pp. 1–13.

[105] C. Wu, J. Liu, X. Wang, and X. Dong, “Chain of reasoning for visual question answering,” in *Proc. Advances Neural Inf. Process. Syst.*, 2018, pp. 275–285.

[106] R. Cadene, H. Ben-younes, M. Cord, and N. Thome, “MUREL: Multimodal relational reasoning for visual question answering,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2019, pp. 1989–1998.

[107] R. Das, S. Dhuliawala, M. Zaheer, and A. McCallum, “Multi-step retriever-reader interaction for scalable open-domain question answering,” in *Proc. 7th Int. Conf. Learn. Representations*, 2019, pp. 1–13.

[108] X. Duan *et al.*, “Watch, reason, and code: Learning to represent videos using program,” in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1543–1551.

[109] W. Yu *et al.*, “Layout-graph reasoning for fashion landmark detection,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2019, pp. 2937–2945.

[110] Y. Xu, L. Qin, X. Liu, J. Xie, and S.-C. Zhu, “A causal and-or graph model for visibility fluent reasoning in tracking interacting objects,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 2178–2187.

[111] H. Xu, C. Jiang, X. Liang, L. Lin, and Z. Li, “Reasoning-RCNN: Unifying adaptive global reasoning into large-scale object detection,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2019, pp. 6412–6421.

[112] M. Narasimhan, S. Lazebnik, and A. Schwing, “Out of the box: Reasoning with graph convolution nets for factual visual question answering,” in *Proc. Advances Neural Inf. Process. Syst.*, 2018, pp. 2654–2665.

[113] P. Xiong, H. Zhan, X. Wang, B. Sinha, and Y. Wu, “Visual query answering by entity-attribute graph matching and reasoning,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2019, pp. 8349–8358.

[114] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, and Y. Kalantidis, “Graph-based global reasoning networks,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2019, pp. 433–442.

[115] X. Liu *et al.*, “Social relation recognition from videos via multi-scale spatial-temporal reasoning,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2019, pp. 3561–3569.

[116] Y.-H. H. Tsai, S. Divvala, L.-P. Morency, R. Salakhutdinov, and A. Farhadi, “Video relationship reasoning using gated spatio-temporal energy graph,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2019 pp. 10416–10425.

[117] Y. Duan, Y. Zheng, J. Lu, J. Zhou, and Q. Tian, “Structural relational reasoning of point clouds,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2019, pp. 949–958.

[118] Y. Wen, Y. Yang, R. Luo, J. Wang, and W. Pan, “Probabilistic recursive reasoning for multi-agent reinforcement learning,” in *Proc. 7th Int. Conf. Learn. Representations*, 2019, pp. 1–19.

[119] X. Chen, L.-J. Li, L. Fei-Fei, and A. Gupta, “Iterative visual reasoning beyond convolutions,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7239–7248.

[120] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar, “Transparency by design: Closing the Gap between performance and interpretability in visual reasoning,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 4942–4950.

[121] Q. Cao, X. Liang, B. Li, G. Li, and L. Lin, “Visual question reasoning on general dependency tree,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7249–7257.

[122] H. Fan and J. Zhou, “Stacked latent attention for multimodal reasoning,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 1072–1080.

[123] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, “Neural module networks,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 39–48.

[124] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the v in VQA matter: Elevating the role of image understanding in visual question answering,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6904–6913.

[125] J. Johnson *et al.*, “CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2901–2910.

[126] D. A. Hudson and C. D. Manning, “GQA: A new dataset for compositional question answering over real-world images,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 6700–6709.

[127] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, “Learning to compose neural networks for question answering,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, Jun. 2016, pp. 1545–1554.

[128] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, “Learning to reason: End-to-end module networks for visual question answering,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 804–813.

[129] J. Johnson *et al.*, “Inferring and executing programs for visual reasoning,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2989–2998.

[130] R. Hu, J. Andreas, T. Darrell, and K. Saenko, “Explainable neural computation via stack neural module networks,” in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 53–69.

[131] G. Li, X. Wang, and W. Zhu, “Perceptual visual reasoning with knowledge propagation,” in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 530–538.

[132] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum, “Neural-symbolic VQA: Disentangling reasoning from vision and language understanding,” in *Proc. Advances Neural Inf. Process. Syst.*, 2018, pp. 1031–1042.

[133] A. S. D. Garcez, K. B. Broda, and D. M. Gabbay, *Neural-Symbolic Learning Systems: Foundations and Applications*. Berlin, Germany: Springer, 2012.

Wenwu Zhu (Fellow, IEEE) received the Ph.D. degree from New York University, New York, NY, USA, in 1996. He is currently a Professor and the Vice Chair of the Department of Computer Science and Technology, Tsinghua University, Beijing, China, the Vice Dean of National Research Center for Information Science and Technology, and the Vice Director of Tsinghua Center for Big Data. Prior to his current post, he was a Senior Researcher and a Research Manager with Microsoft Research Asia. He was the Chief Scientist and the Director with Intel Research China from 2004 to 2008. He worked with Bell Labs New Jersey as a Member of Technical Staff during 1996–1999. His current research interests are in the area of data-driven multimedia networking and cross-media big data computing. He has authored or coauthored more than 350 referred papers, and is inventor or co-inventor of more than 50 patents. He received eight Best Paper Awards, including the ACM Multimedia 2012 and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY in 2001 and 2019. He served as an Editor-in-Chief for the IEEE TRANSACTIONS ON MULTIMEDIA. He served in the steering committee for IEEE TRANSACTIONS ON MULTIMEDIA (2015–2016) and IEEE TRANSACTIONS ON MOBILE COMPUTING (2007–2010). He serves as a General Co-Chair for ACM Multimedia 2018 and ACM CIKM 2019. He is an AAAS Fellow, SPIE Fellow, and a member of The Academy of Europe (Academia Europaea).

Xin Wang (Member, IEEE) received the B.E. and Ph.D. degrees in computer science and technology from Zhejiang University, Hangzhou, China, and the Ph.D. degree in computing science from Simon Fraser University, Burnaby, BC, Canada. He is currently a Research Assistant Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include cross-modal multimedia intelligence and inferable recommendation in social media. He has authored or coauthored several high-quality research papers in top conferences including ICML, MM, KDD, WWW, SIGIR, etc. He was the recipient of 2017 China Postdoctoral Innovative Talents Supporting Program.

Wen Gao (Fellow, IEEE) received the Ph.D. degree in electronics engineering from The University of Tokyo, Tokyo, Japan, in 1991. He is currently a Boya Chair Professor, and the Dean with the School of Electronics Engineering and Computer Science, Peking University, Beijing, China. He is the Director of Pengcheng Laboratory, Shenzhen, China. He is the President of Chinese Computer Federation since February 2016. He was the Vice President of National Natural Science Foundation of China from February 2013 to February 2018. He joined the Harbin Institute of Technology from 1991 to 1995, as a Professor, the Department Head of Computer Science. He was with the Institute of Computing Technology, Chinese Academy of Sciences, from 1996 to 2005. He has been with the Peking University as a Professor since 2006. He works in the areas of multimedia and computer vision, including video coding, video analysis, multimedia retrieval, face recognition, multimodal interfaces, and virtual reality. His most cited contributions are model-based video coding and face recognition. He has authored or coauthored 6 books and more than 1000 technical articles in refereed journals and proceedings in the above areas. He earned many awards including six State Awards in Science and Technology Achievements. He has been featured by IEEE Spectrum in June 2005 as one of the "Ten-To-Watch" among China's leading technologists. He is a fellow of ACM and a member of the Chinese Academy of Engineering.