Textual-Visual Reference-Aware Attention Network for Visual Dialog

Dan Guo, Hui Wang, Shuhui Wang, and Meng Wang, *Senior Member, IEEE*

***Abstract*—Visual dialog is a challenging task in multimedia understanding, which requires the dialog agent to answer a series of questions that are based on an input image. The critical issue to produce an exact answer is how to model the mutual semantic interaction among feature representations of the image, question-answer history, and current question. In this study, we propose a textual-visual Reference-Aware Attention Network (RAA-Net), which aims to effectively fuse Q (question), H (history), V_l (local vision), and V_g (global vision) to infer the exact answer. In the multimodal feature flows, RAA-Net first learns the textual context through multi-head attention between Q and H and then guides the textual reference semantics to the image to capture visual reference semantics by self- and cross-reference-aware attention in and between V_l and V_g. In the proposed RAA-Net, we exploit the two-stage (intra- and inter-) visual reasoning mechanism on V_l and V_g. Extensive experiments on the VisDial v0.9 and v1.0 datasets show that RAA-Net achieves state-of-the-art performance. Visualization results on both visual and textual attention maps further validate the remarkable interpretability achieved by our solution.**

***Index Terms*—Visual dialog, attention network, textual reference, visual reference, multimodal semantic interaction.**

I. Introduction

RECENTLY, cross-modal semantic understanding [1]–[5] between vision and language [1]–[3] has received considerable attention, such as image captioning [6]–[11], visual grounding [12]–[17], and visual question answering (VQA) [18]–[22]. In these studies, semantic referring between vision and language is performed in a one-way single round. Taking VQA as an example, the agent is required to first understand a specific question, then ground the relevant visual contents in the image, and finally infer an answer. In contrast, visual dialog [23]–[26], as an extension of VQA, is bidirectional under multi-round question-answer (QA) pairs discussing the same image. It has a stronger correlation in the semantic feedback during dialog. In the visual dialog task, grasping sufficient semantic relationships among multi-round multi-modality data for accurate answer reasoning is essential.

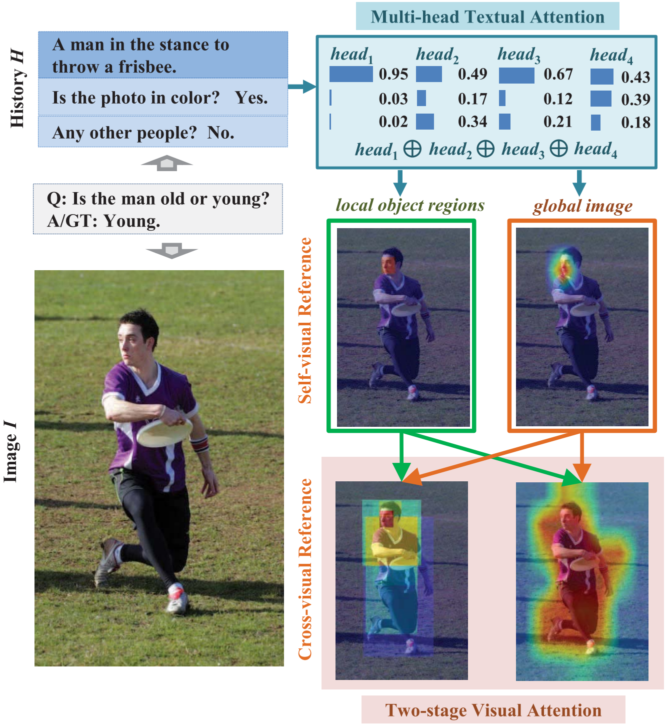

Fig. 1. Illustration of textual and visual attention maps in our solution on the basis of an image sample in VisDial v0.9. The use of multi-modality co-reference helps to enhance the correct semantic inference. Two-stage visual referring between local and global visions are guided by the textual reference between question and history.

Early dialog studies addressed multimodal semantic reasoning based on the feature fusion of vision and language in a single-step, which is similar to VQA [18]. The state-of-the-art approaches focus more on various co-attentions on feature representations of question Q, history H, and answer A, yielding many promising results [27]–[29]. These approaches mostly involve global image-based features. In our paper, the complementarity between local and global visions provides richer visual semantics compared with a single type of visual information. Modeling the mutual contextual correlation among them is beneficial to perform accurate visual grounding.

In this study, we attempt to comprehensively understand the image content and capture the latent context relationships in the question and history more deeply. As illustrated in Fig. 1,

Manuscript received September 20, 2019; revised February 7, 2020 and March 15, 2020; accepted April 7, 2020. Date of publication May 18, 2020; date of current version July 6, 2020. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61725203, Grant 61732008, Grant 61672497, and Grant 61876058. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christophoros Nikou. *(Corresponding author: Meng Wang.)*

Dan Guo, Hui Wang, and Meng Wang are with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China (e-mail: guodan@hfut.edu.cn; wanghui.hfut@gmail.com; eric.mengwang@gmail.com).

Shuhui Wang is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wangshuhui@ict.ac.cn).

Digital Object Identifier 10.1109/TIP.2020.2992888

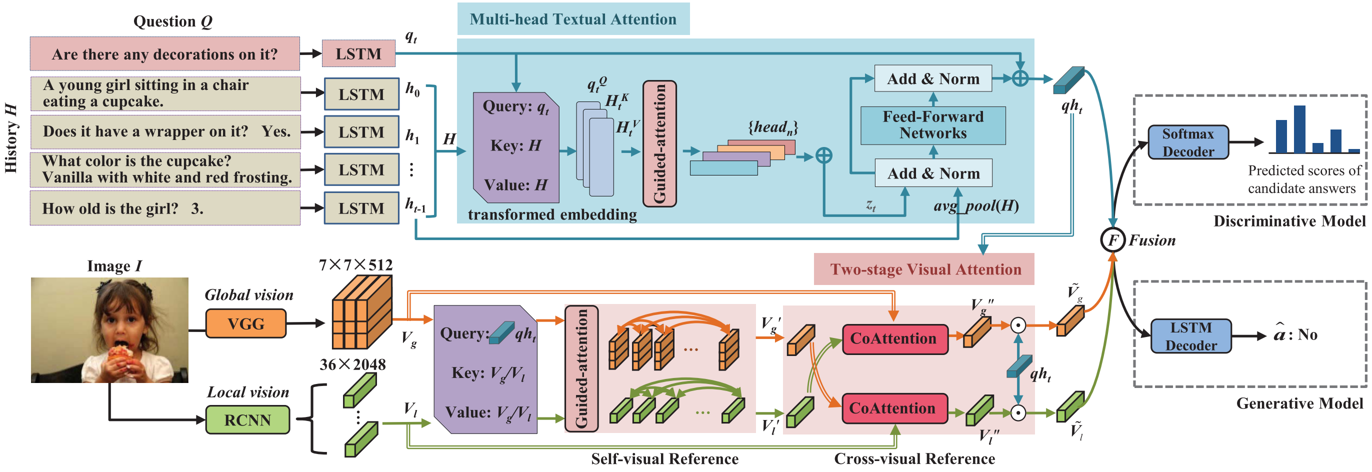

Fig. 2. Overall framework of textual-visual RAA-Net for visual dialog.

we address the problem of textual-visual semantic correlation from the following two aspects. First, with respect to the current question Q: "Is the man old or young?", the textual reference process recalls the word "man" in history snippets. The history sentence "A man in the stance to throw a frisbee." corresponds to a high attentive concentration, while the other two QA pairs with somewhat unrelated textual contents do not. Second, subsequent to reviewing textual cues, we exploit the visual reference using a two-step visual reasoning. In the 1st-stage visual reasoning using self-visual reference-aware attention, both local and global attention maps consistently focus on the man's face. The question inference is complicated and requires more context cues. Then, in the 2nd-stage visual reasoning using cross-visual reference-aware attention, both local and global attention maps are expanded to cover the whole body of the man. These attended visual semantics help obtain an exact answer.

Fig. 2 illustrates the entire framework in this study. We propose a Reference-Aware Attention Network (RAA-Net), which progressively tackles textual and visual references that are conditioned on the current question. The core idea obeys the pipeline of the semantic interaction in Fig. 1. To imitate the way that humans solve question Q, in RAA-Net, we first explore multi-head textual attention that **reviews** history H multiple times under the instruction of Q. Then, we design two-stage visual attention that **re-attends** the visual cues of both local objects and global image under the joint guidance of Q and H.

Specifically, RAA-Net accomplishes the sentence-level textual attention and associates the textual semantics of $[Q; H]$ to discover the relevant visual cues. In other words, once obtaining the textual co-reference between Q and H, RAA-Net devotes to visual co-reference learning of different visual contents in image I, which utilizes both local object-wise features and global feature maps. Different from that "reviewing" textual contents multiple times, there are merely two alternate "glimpses" at image I. The first "glimpse" captures each intra-visual cues by guided-attention, and the second "glimpse" explores the inter-visual correlation using co-attention. These two types of attention mechanisms can discover subtle visual cues more comprehensively. Finally, an answer is inferred by a multimodal semantic fusion module. In a nutshell, RAA-Net enhances the salient semantics in rich context acquisition by cross-modal correlation learning; therefore, promising performances can be achieved.

The main contributions are summarized as follows:

- In the proposed RAA-Net, textual-visual semantic reasoning is conducted by multimodal reference-aware attention learning. We gradually capture the question-conditioned cues from history, global and local visions.
- RAA-Net consists of multi-head textual and two-stage visual attention modules. The former excels at potential textual mining between question and history, and the latter performs well on intra- (self-) and inter- (cross-) visual references. Both of them are beneficial in identifying the true attentive semantics within a modality and learning the visual-linguistic correlations.
- Experimental results on the VisDial v0.9 and v1.0 datasets show that the proposed approach achieves promising performance compared to the state-of-the-art approaches.

The remainder of this paper is organized as follows. Section II reviews related works. Section III elaborates on the proposed RAA-Net model. The analysis and discussion of the experimental results are presented in Section IV, and conclusions are given in Section V.

II. Related Work

This section reviews studies that are related to three aspects as follows: visual dialog, attention-based models and dual visual correlation studies.

A. Visual Dialog

Visual dialog is a new iterative visual-language task, which is taken as an extension of VQA. Recently, two popular dialog datasets have been introduced by [23] and [24]. De Vries *et al.* [23] have collected a GuessWhat?! dataset via a two-player image-guessing game. Given an image and a caption description, one player asked questions to guess what objects appeared in the image, while the other player answered yes/no/NA. However, the questions in the dialog are closed-ended. Das *et al.* [24] have released the largest visual dialog dataset VisDial, which paired two annotators on Amazon Mechanical Turk to collect more free-form questions and answers. The open-ended question types include number, color, weather, *etc.*, and the answer can be yes/no or even a more elaborate description.

Most of the proposed approaches for visual dialog are based on encoder-decoder framework, and current encoder-based studies can be divided into two categories. (1) **Fusion-based models.** LF (Late Fusion) [24] encoded the question, dialog history, and image separately and then concatenated them to a joint representation for answer inferring. HRE (Hierarchical Recurrent Network) [24] first used LSTM to encode the joint feature of question and image, and another LSTM to encode each QA pair in the dialog history. Finally, dialog-level LSTM was applied to capture the temporal correlation in the whole dialog, and the output was used to decode the final answer. Both the LF and HRE methods are introduced in [24], but fused the multimodal features at different stages. The former focused on the joint multimodal feature learning, whereas the latter adopted a hierarchical recurrent network for history encoding. (2) **Graph-based models.** More recently, some studies have been proposed that are based on various graph neural networks (GNN). As the special structure of graphs, these methods are very suitable for reasoning-style tasks. Zheng *et al.* [30] have proposed an EM-style GNN, which constructed the graph on the whole dialog and applied EM algorithm to infer the answer based on the current question and QA history. Schwartz *et al.* [31] have proposed a factor graph attention mechanism, which constructed the graph overall the multimodal features and estimated their attention interactions.

B. Attention-Based Model

To further improve performance, various attention mechanisms have been widely used in visual dialog. In [27], Lu *et. al.* have proposed a novel history-conditioned image attention model (HCIAE), which attended to image features according to the dialog context. Wu *et al.* [28] have introduced a sequential co-attention model (CoAtt) and trained the model with the reinforcement learning strategy. Furthermore, there are some studies that focus on locating the pronouns or nouns (in languages) with visual objects (in image), which deemed as the visual reference resolution. Seo *et al.* [32] have designed an attention-based memory module named AMEM to store previous visual attention maps; the current question used the memory to solve the visual reference issue. Kottur *et al.* [33] have imposed the attention mechanism on a set of neural module networks [34] to handle visual reference resolution at the word-level.

All works mentioned above focused on visual reference or multimodal co-attention learning. The visual dialog is an iterative textual-visual interaction process. An excellent textual attention mechanism is also essential. Vaswani *et al.* [35] have introduced a scaled dot-product attention mechanism, which calculated the scaled similarity matrix of the input $query$ with all $keys$, and applied a softmax function to obtain the weights on the input $values$. On the basis of this attention mechanism, Vaswani *et al.* have introduced a transformer framework for machine translation in the field of natural language processing. Without traditional CNN and RNN operations, the transformer produced better performance by parallelism attention computing on the input sentence. The novel scaled dot-product attention and multi-head attention in a transformer inspired many studies [36], [37], which indicates that the transformer method is not limited to machine translation. Inspired by [35], we design a multi-head textual attention module to address the textual-reference problem. Different from the totally textual $query$, key, and $value$ in the traditional transformer methods, in our solution, we propose a multimodal scaled dot-product attention mechanism as follows. For the textual-reference process, scaled dot-product attention is first used to attend strongly related history snippets that are conditioned on the current question. Then, for the visual-reference process, a variant of the scaled dot-product attention is implemented to focus on related visual regions under the guidance of the textual semantics.

C. Dual Visual Correlation

Different visual representations involving both local object-based and global image-based visual features have been introduced for VQA [38], [39]. Both these two works have co-attended to question, local and global visions. In [38], a hierarchical co-attention scheme has been proposed; the local and global visual features were attended respectively, and they were fused to infer the answer. Lu *et al.* [39] have employed a multiplicative embedding scheme to jointly attend the question-conditioned global image area and local objects. The method in [39] is most closely related to ours. Lu *et al.* have applied a dual visual attention correlation, which is similar to the visual-reference process in our model. However, the work [39] was merely performed on original visual features at once. In contrast, we adopt a progressive attention mechanism to obtain fine-grained visual representations. We design a two-stage visual attention, in which the 1st-stage is used to capture respective self-visual cues using guided-attention, and the 2nd-stage explores the visual cue correlation using co-attention.

III. Proposed Method

A. Problem Definition

The visual dialog task [24] is defined as follows: the input consists of an image I and (t-1)-round dialog-history $H = (c, (q_1, a_1), \ldots, (q_{t-1}, a_{t-1}))$, where c is the image caption, and (q, a) is the previous QA pair. To answer the question Q at the current round t, the agent can decode in either of the following two tactics: discriminative and generative models. Discriminative models select the answer with the maximum score from a list of N candidate answers $A_t = \{a_t^{(1)}, \ldots, a_t^{(N)}\}$, while generative models decode answers by sequential learning. The decoding process of generative models is optimized by maximizing the log-likelihood of the ground-truth answer sequence $a_t^{gt} \in A_t$.

B. Method Overview

1) Method Overview: As a vision-language interaction task, the visual dialog can be divided into three aspects: textual comprehension of question Q and history H, visual grounding of image I, and answer inference for question Q. As illustrated in Fig. 2, the proposed model RAA-Net includes three corresponding modules: (1) question-conditioned multi-head textual attention for textual comprehension in Section III-C, (2) two-stage visual attention for correlated visual and textual component association in Section III-D, and (3) multimodal semantic fusion for answer inference in Section III-E.

In this study, we aim to achieve effective multimodal semantic understanding of question Q, history H, local vision V_l, and global vision V_g to inferring the answer A. The proposed RAA-Net model is a query-adaptive reasoning network. It first explores textual attention, which applies the question to pick out related history snippet semantics and output the textual reference-aware representation qh_t. Next, it implements visual attention by imposing qh_t on V_l or V_g to discover local/global intrinsic visual cues V_l'/V_g', namely self-visual reference within each single modality. Then, it performs a cross-visual reference. The visual cues V_l' and V_g' are utilized cross-wise to realize mutual visual correlation, *i.e.*, $(qh_t + V_g' + V_l) \rightarrow V_l''$ and $(qh_t + V_l' + V_g) \rightarrow V_g''$. In other words, RAA-Net first performs textual inferring, and then independently and mutually addresses local and global visual correlation inferring.

2) Attention Backbone: To build the reference-aware attention network RAA-Net, we adopt two types of attention mechanisms, *i.e.*, guided-attention and co-attention. The former measures intra-relationship in a feature sequence X guided by a query feature Y, while the latter integrates inter-relationship among multiple features, *e.g.*, $(X, X_1, X_2, \cdots, X_n)$.

Guided-attention: We apply this attention to explore the interaction in a single feature sequence. Then the relationship within the feature sequence (*i.e.*, attention distribution) is used to learn a new embedding of the sequence. Motivated by scaled dot-product attention [35], in our solution, the guided-attention is defined as follows. Given a feature sequence $X \in \mathbb{R}^{l\times d}$ and a query feature $Y \in \mathbb{R}^{1\times d}$, three learnable parameters $W^Q, W^K, W^V \in \mathbb{R}^{d\times\lambda}$ project the inputs to three new feature sequences: ***query*** $Y^Q \in \mathbb{R}^{1\times\lambda}$, ***key*** $X^K \in \mathbb{R}^{l\times\lambda}$, and ***value*** $X^V \in \mathbb{R}^{l\times\lambda}$, where l is the length of the input sequence X. The attention function is performed as shown in Eq. (1). Let $Att_{guide} \in \mathbb{R}^{1\times l}$ be the output of the attention, which is the weighted sum of value X^V based on the scaled similarity matrix $[\frac{Y^Q(X^K)^T}{\sqrt{\lambda}}]$. It handles the new embedding for the feature sequence X under Y as follows:

$$Att_{guide}(Y^Q, X^K, X^V) = softmax(\frac{Y^Q(X^K)^T}{\sqrt{\lambda}}) \cdot X^V \quad (1)$$

where λ is the transformed dimension parameter which is used to balance the attention distribution, $(\cdot)^T$ denotes matrix transposition, and $softmax(\frac{Y^Q(X^K)^T}{\sqrt{\lambda}})$ calculates the intra-attention weighting distribution of X.

Co-attention: Different from the above mentioned intrinsic correlation in a feature sequence X, we design a co-attention to weigh the influences of other feature sequences on X. The effect of the combination of $\{X_1, X_2, \cdots, X_n\}$ on X is formulated as follows:

$$\begin{cases} \mu = tanh((X_1W_1 + X_2W_2 + \ldots + X_nW_n)\mathbb{1}^T + XW) \\ Att_{co}(X, \{X_1, \ldots, X_n\}) = softmax(\mu W_{co}) \end{cases} \quad (2)$$

where W, W_i ($i \in [1, n]$), and W_{co} are learnable parameters that project $\{X, X_1, \cdots, X_n\}$ into the same feature dimension, and $\mathbb{1}^T \in \mathbb{R}^l$ is a vector with all elements set to 1. Here, we use non-linearity $tanh$, instead of classical $sigmod$, to compute μ, where $tanh$ squashes the input to the range of [-1, 1]; μW_{co} is later applied to $softmax$ with the range of (0, 1). Compared with $sigmoid'$, $tanh'$ has larger gradient range results, which results in faster convergence and better inhibition of gradient vanishing [40]. Moreover, $tanh \in [-1, 1]$ is zero-symmetric, while $sigmod \in (0, 1)$. If the inputting data is always greater than zero, this will lead to the offset phenomenon. Considering these factors, we choose $tanh$ for computing μ.

C. Textual Reference-Aware Attention

Visual dialog refers to a multi-round conversation about the image. Questions in a dialog always contain at least one pronoun (*e.g.*, "he", "she", "it", and "this"). The latent semantic co-reference between question Q and history H is still a challenge to be solved. In our framework, we propose multi-head textual attention for utilizing Q and H, which is an extended guided-attention to address textual inferring.

1) Transformed Textual Embedding: Prior to the textual reference process, we obtain sentence-level features of Q and H using respective LSTM [41]; we denote them as $LSTM_Q$ and $LSTM_H$. Each word x_i in the sentence is assigned to a one-hot vector and modeled by a learnable word embedding matrix W_e. As for question Q, we take the last hidden state $LSTM_Q(x_L W_e)$ as the question feature q_t, where L is the number of words in question Q. We adopt $LSTM_H$ to encode the history feature $H = [h_0, h_1, \ldots, h_{t-1}]$ in the same way as that of Q, where each QA pair is taken as a whole sentence and h_0 denotes the textual feature of image caption c. Here, $H \in \mathbb{R}^{t\times d_t}$, where d_t is the dimension of these textual features.

In this study, we aim to track the context feedback from a sequence {caption c and each QA pair} in H under the guidance of question Q. It is advantageous to adopt the guided-attention in Eq. (1). Thus, we transform original textual features into new features *query* q^Q, *key* H^K, and *value* H^V as follows:

$$\begin{cases} q^Q = Linear(q_t; W^Q) \\ H^K = Linear(H; W^K) \\ H^V = Linear(H; W^V) \end{cases} \quad (3)$$

where W^Q, W^K and W^V are learnable parameters and $Linear(\cdot)$ denotes a fully connected layer.

2) Multi-Head Textual Attention: Up to now, we have obtained new query semantics of question Q (*query* q^Q), to-be-attended embedding (*key* H^K) and new projection (*value* H^V) of history H. The operation $q^Q(H^K)^T$ reflects the attention distribution of H under Q. With the input $\{q^Q, H^K, H^V\}$, we use the guided-attention to extract relevant textual semantics from the dialog history.

Textual semantic reasoning in natural languages is usually complicated. The use of only a single-step guided-attention to a multi-round conversation may be insufficient. Thus, we design a multi-head guided-attention, which imitates the process that humans use to review the conversational interaction multiple times. The inputs of the multi-head attention module are q^Q, H^K, and H^V. At head h, the attention is defined as $A_h = Att_{guide}(q^Q, H^K, H^V)|_h \in \mathbb{R}^{1\times t}$ by Eq. (1). Thus, the textual reference scheme $Ref_{T:Q\leftrightarrow H}$ is calculated as follows:

$$
\begin{aligned}
&Ref_{T:Q\leftrightarrow H}:\\
&\quad z_t = Linear\Big(Multihead(q^Q, H^K, H^V); W^z\Big)\\
&\quad\quad = Linear\Big(\big[Att_{guide}(q^Q, H^K, H^V)|_{head=1}, \cdots\\
&\quad\quad\quad Att_{guide}(q^Q, H^K, H^V)|_{head=h}\big]; W^z\Big)\\
&\quad\quad = Linear\Big(\big[A_1 \oplus A_2 \oplus \ldots \oplus A_h\big]; W^z\Big) \qquad (4)
\end{aligned}
$$

where h is the number of heads, and $\oplus$ is the concatenation operation.

Currently, z_t denotes the question-conditioned textual semantics of H by the multi-head attention. In addition, we add z_t with the average pooling of H. $Avg\_Pool(H)$ can be regarded as a complement of z_t. z_t is a local, fine-grained, historical relevance calculation, whereas $Avg\_Pool(H)$ considers the original global historical semantics. The joint exploitation of $Avg\_Pool(H)$ and z_t is proposed to learn textual semantics much more comprehensively. After the two layer normalization [42] and one feed-forward network, we obtain a question-conditioned history feature $\hat{h}_t$ as follows:

$$
\begin{cases}
\hat{z}_t = LayerNorm(z_t + Avg\_Pool(H))\\
h_t = Linear\Big(ReLU(\hat{z}_t W_1^z); W_2^z\Big)\\
\hat{h}_t = LayerNorm(h_t + \hat{z}_t)
\end{cases} \qquad (5)
$$

where W_1^z and W_2^z are learnable parameters.

We concatenate the question feature q_t and the new history feature $\hat{h}_t$ to generate the textual reference-aware vector qh_t.

$$
qh_t = [q_t \oplus \hat{h}_t] \qquad (6)
$$

D. Visual Reference-Aware Attention

Subsequent to textual reference, we conduct two-stage visual reference. As illustrated in Fig. 2, the proposed RAA-Net includes both intra- and inter- visual inferring, *i.e.*, respective self-attention of local and global visual features, and cross-reference (co-attention) between them.

1) Transformed Visual Embedding: Here, we use both local object-based and global image-based visual features to achieve comprehensive visual inferring. Faster R-CNN [21], [43] pretrained on Visual Genome dataset [44] is applied to to extract local object-based features $V_l \in \mathbb{R}^{36\times 2048}$, where 36 is the number of detected objects for each image, and 2048 is the dimension of local visual features. Besides, we adopt a pretrained model VGG19 [45] to extract global visual features. The output of the last pooling layer of VGG19 is denoted as $V_g \in \mathbb{R}^{7\times 7\times 512}$, where 7×7 is the spatial size and 512 is the channel number of the feature maps.

To perform better self-visual reasoning, we adopt the same function of transformed embedding in Eq. 3, which targets to obtain more effective representations ($query$, key and $value$). The difference is that here it is a multimodal transformed embedding. We use the textual reference-aware vector qh_t to build $query$, which remains the joint textual co-reference semantics from Q and H. For either visual features V_l or V_g, we transform them into respective new embedding of *key* and *value*. The original qh_t, V_g, and V_l are mapped into each new embedding space as follows:

$$
\begin{cases}
q_l^Q = Linear(qh_t; W^{Q_{q_l}})\\
V_l^K = Linear(V_l; W^{K_{V_l}})\\
V_l^V = Linear(V_l; W^{V_{V_l}})
\end{cases}
\qquad
\begin{cases}
q_g^Q = Linear(qh_t; W^{Q_{q_g}})\\
V_g^K = Linear(V_g; W^{K_{V_g}})\\
V_g^V = Linear(V_g; W^{V_{V_g}})
\end{cases} \qquad (7)
$$

where $q_{l|g}^Q \in \mathbb{R}^{1\times d}$, $V_g^{K|V} \in \mathbb{R}^{M\times d}$, $V_l^{K|V} \in \mathbb{R}^{K\times d}$, $K = 36$, and $M = 7\times 7$. $\{W^{Q_{q_l}}, W^{K_{V_l}}, W^{V_{V_l}}\}$ and $\{W^{Q_{q_g}}, W^{K_{V_g}}, W^{V_{V_g}}\}$ are all learnable parameters.

2) 1st-Stage Visual Attention (Self-visual Reference): This stage targets to identify the most related visual cues within each data space. Either based on local or global visual features, we implement the guided-attention under the textual reference-aware semantics qh_t. In other words, this stage implements the reference $(qh_t + V_l) \rightarrow V_l'$ and $(qh_t + V_g) \rightarrow V_g'$, where V_l' and V_g' denote the intra- local and global visual cues.

Taking local visual features as an example, given the **textual *query*** q_l^Q, **visual *key*** and ***value*** (V_l^K and V_l^V), $V_l{}'$ is obtained by weighted summation over all feature vectors in V_l^V with respect to q_l^Q and V_l^K. Similarly, we calculate the intra-global visual cue $V_g{}'$ in the same way. Thus, we perform the visual grounding of related objects and salient relevant regions in their self-visual view, respectively. Essentially, it is a self-visual reference process.

Step 1: $Ref_{V_l{}':V_l\leftrightarrow V_l|QH}$:

$$
V_l{}' = Att_{guide}(q_l^Q, V_l^K, V_l^V) \in \mathbb{R}^{1\times d} \qquad (8)
$$

Step 2: $Ref_{V_g{}':V_g\leftrightarrow V_g|QH}$:

$$
V_g{}' = Att_{guide}(q_g^Q, V_g^K, V_g^V) \in \mathbb{R}^{1\times d} \qquad (9)
$$

Notably, we set $\lambda = 1$ in Eq. (8) and Eq. (9). The effect of λ with a high value is to make the calculated attention distribution more scattered and smooth [35]. $\lambda = 1$ without scaling the attention distribution is suitable to a concentrated visual attention.

3) 2nd-stage Visual Attention (Cross-visual Reference): Different from the previous stage utilizing visual cues in each visual view (*i.e.*, $V_l \leftrightarrow V_l, V_g \leftrightarrow V_g$), this stage exploits the mutual correlation (*i.e.*, $V_l \leftrightarrow V_g$). As shown in Fig. 1, the cross-visual reference captures the surrounding spatial context to realize local-global visual correlation, which makes the attentive areas of both sides to become more and more consistent and accurate. We adopt the co-attention to implement the mutual correlation between V_g and V_l. To be specific, it implements $(qh_t + V_g' + V_l) \rightarrow V_l''$ for local vision and $(qh_t + V_l' + V_g) \rightarrow V_g''$ for global vision, where V_l'' and V_g'' denote the outputs of local and global reference-aware visual cues. V_l'' and V_g'' are performed as follows:

Step 3: $Ref_{V_l{''}:V_l \leftrightarrow V_g|QH}$:

$$\begin{cases} \alpha_l = Att_{co}\Big(V_l, \{V_g', qh_t\}\Big) \in \mathbb{R}^{1\times K} \\ V_l'' = \alpha_l V_l \end{cases} \tag{10}$$

Step 4: $Ref_{V_g{''}:V_l \leftrightarrow V_g|QH}$:

$$\begin{cases} \alpha_g = Att_{co}\Big(V_g, \{V_l', qh_t\}\Big) \in \mathbb{R}^{1\times M} \\ V_g'' = \alpha_g V_g \end{cases} \tag{11}$$

Up to now, visual reference-aware representations V_l'' and V_g'' have already be explored under the guidance of the textual-aware semantics qh_t. As the Hadamard product $\odot$ (*i.e.*, element-wise multiplication) has been confirmed to be effective at enhancing the textual-visual correlation in VQA [46], here we use it to refine the visual features.

$$\begin{cases} \tilde{V}_l &= V_l{''} \odot f(qh_t) \in \mathbb{R}^{1\times d} \\ \tilde{V}_g &= V_g{''} \odot f(qh_t) \in \mathbb{R}^{1\times d} \end{cases} \tag{12}$$

where $f(\cdot)$ denotes a non-linear transformation function as follows:

$$f(x) = tanh(xW_{f1}) \odot \sigma(xW_{f2}) \tag{13}$$

where W_{f1} and W_{f2} are learnable parameters.

E. Multimodal Semantic Fusion

Finally, we perform a joint semantic embedding e_t, which fuses the above mentioned reference-aware features including the textual feature qh_t, the local and global visual features $\tilde{V}_l$ and $\tilde{V}_g$.

$$e_t = tanh\Big(Linear\big([qh_t \oplus \tilde{V}_l \oplus \tilde{V}_g]; W^f\big)\Big) \tag{14}$$

where $W^f \in \mathbb{R}^{4d\times d}$ is a learnable parameter. As for the subsequent decoding setting, we obey the rule in [24]. In the generative model, e_t is fed into a LSTM-based decoder to infer the answer $\hat{a}$; in the discriminative model, e_t is input into the softmax decoder to sort the candidate answers A_t. The training details are in Section IV-A.3.

IV. Experiments

A. Experimental Setup

1) Datasets: The experiments are conducted on the VisDial v0.9 and v1.0 datasets [24]. VisDial v0.9 is collected through a two-player image-guessing game that is based on COCO [47] images. For each image example, the dialog consists of 10-round QA pairs. VisDial v0.9 contains 83k and 40k dialogs on COCO-train and COCO-val images, respectively, with a total of 1.2M QA pairs. VisDial v1.0 is an extension of VisDial v0.9, which adds additional 10k dialogs on Flickr images. The new train, validation, and test splits have 123k, 2k, and 8k dialogs, respectively. It is worth noting that in the test split of VisDial v1.0, each dialog has flexible m rounds of QA pairs, where m is in the range of 1 to 10.

2) Evaluation Metrics: Following [24], the answer accuracy is evaluated by retrieving the ground-truth answer from a list of 100 option answers. We adopt the following retrieval metrics: (1) average rank of the ground-truth answer (**Mean**), (2) recall rate of the ground-truth answer in top-k ranked option answers (**R@k**), and (3) mean reciprocal rank of the ground-truth answer (**MRR**). In addition, VisDial v1.0 introduces an additional retrieval metric normalized discounted cumulative gain (**NDCG**), which penalizes the lower rank of answers with high relevance.

3) Implementation Details: We build the vocabulary containing the words occurring at least five times in the training split. The lengths of captions, questions, and answers are truncated to 40/20/20 for discriminative models, and to 24/16/8 for generative models, respectively. Words in all captions, questions, and answers are embedded into 300-dim vectors by the GloVe embedding [48]. For the local visual feature extraction, we adopt the implementation of object detection using Faster R-CNN for VQA [24], [46], including choosing $K = 36$ detected objects. In addition, in Eqs. 1∼14, except for the parameterless activation functions *softmax, tanh, ReLU,* and *Avg_Pool* (mean pooling), only the *Linear* (*i.e.*, *FC*, the fully connected layer) remains. There are many FC operations, as our task is a cross-modality problem. FC is used as a conventional projection operation to project different features (textual and visual semantics) into a much more close embedding space, promising better correlation learning and relation reasoning.

In our implementation, the Adam optimizer [49] is initialized with the learning rate of 4×10^{-4} and multiplied by 0.5 after every 10 epochs. There are three LSTM modules. We denoted them as $LSTM_Q$, $LSTM_H$, and $LSTM_A$, which extract the features of question Q, history H, and answer A, respectively. They are independently initialized and trained without sharing parameters. All of LSTMs are set with 1-layer and 512 hidden states. We set the dropout [50] ratio to 0.1 for all attention modules and to 0.5 for the multimodal semantic fusion module. Finally, the generative model is trained with the MLE (Maximum Likelihood Estimation) loss, while the discriminative model is trained with a multi-class N-pair loss [27].

B. Ablation Study of RAA-Net

1) Multi-Head Settings in Textual Reference: We conduct ablative experiments on the validation set of VisDial v0.9. At first, we test different settings of multi-head textual attention. We set the dimension of the textual reference-aware vector to be 512 and test the influence of the head number h

TABLE I

ABLATION STUDY OF DISCRIMINATIVE MODELS WITH DIFFERENT MULTI-HEAD SETTINGS ON VISDIAL VAL V0.9

Model	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
1 head / 512-dim	0.6635	53.28	82.60	90.58	3.97
2 head / 256-dim	0.6641	53.29	82.66	90.66	3.95
4 head / 128-dim	0.6659	53.49	82.79	90.74	3.92
8 head / 64-dim	0.6644	53.35	82.64	90.56	3.97
4 head / 128-dim	0.6659	53.49	82.79	90.74	3.92
4 head / 256-dim	**0.6683**	**53.80**	**82.99**	**90.86**	**3.89**
4 head / 512-dim	0.6667	53.66	82.83	90.78	3.92

TABLE II

ABLATION STUDY OF DISCRIMINATIVE MODELS WITH DIFFERENT TEXTUAL FEATURES ON VISDIAL VAL V0.9

	Discriminative Models				
	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
RAA-Bert	0.5821	44.38	74.82	85.14	5.64
RAA w/o $Avg(H)$	0.6652	53.51	82.61	90.68	3.96
RAA-Net	**0.6683**	**53.80**	**82.99**	**90.86**	**3.89**

under $h = 1$, 2, 4, and 8. With the same dimension $d' = 512$, the dimension of each head is divided as $d_h = d'/h$, *i.e.*, 512, 256, 128, and 64 respectively. As shown in Table I, when the number of heads is $h = 4$, our model gets better performance. We deem each head attention imitating human to review the textual reference at one time. The experimental results show that "reviewing" too many times may disturb the correct reference process, while reviewing fewer times may miss some important textual cues. And $h = 4$ is practical for the task. Then, we test the head dimension d_h. We set the same head number $h = 4$ and test different $d_h = 128$, 256, and 512. RAA-Net achieves the best performance with $d_h = 256$. Thus, we set $h = 4$ and $d_h = 256$ in the following experiments. Table I shows that the performance of RAA-Net at different settings is very similar. This result confirms that RAA-Net has good robustness on the multi-head textual attention.

2) Textual Features: Here, we validate different textual features in the textual reference process. We experiment with **RAA-Bert** and **RAA w/o** $Avg(H)$. RAA-Bert is a variant of RAA-Net, which merely replaces the original sentence-level LSTM features by the mean pooling in the word embedding features extracted by a pre-trained BERT [51]. RAA w/o $Avg(H)$ is a variant of RAA-Net that adopts only z_t to obtain the question-conditioned history features $\hat{h}_t$ (*i.e.*, $\hat{h}_t = z_t$). As shown in the following Table II, RAA-Bert has a noticeable performance drop. There are different usages of the self-attention. The self-attention in BERT [51] explores the word-level relationship for textual embedding, whereas the self-attention in our solution emphasizes correlation learning of multiple multimodal features at sentence level. In addition, compared with RAA w/o $Avg(H)$, RAA-Net has a slight performance improvement. It indicates that z_t contributes to the primary historical influence. Owing to the better performance of RAA-Net, we retain $Avg(H)$ in the proposed RAA-Net.

3) Variants of RAA-Net: To verify each component in RAA-Net, we propose few variants for ablation study:

TABLE III

ABLATION STUDY OF DISCRIMINATIVE MODELS ON VISDIAL VAL V0.9

	Discriminative Models				
	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
RAA w/o V_l	0.6316	49.27	80.01	88.57	4.49
RAA w/o V_g	0.6574	52.64	81.84	90.07	4.14
RAA w/o T-att	0.6605	52.92	82.31	90.38	4.02
RAA w/o V-self-att	0.6631	53.30	82.39	90.44	4.01
RAA w/o T-V-self-att	0.6584	52.76	82.13	90.21	4.14
RAA w/o V-cross-att	0.6404	50.48	80.68	89.22	4.34
RAA-Net	**0.6683**	**53.80**	**82.99**	**90.86**	**3.89**

- **RAA w/o** V_l denotes the variant of RAA-Net with only global visual feature V_g and textual feature qh_t.
- **RAA w/o** V_g denotes the variant of RAA-Net with only local visual feature V_l and textual feature qh_t.
- **RAA w/o T-att** means that under local and global visual feature (V_l and V_g), RAA-Net removes the multi-head textual attention. Instead, the average pooling for history features is calculated as $\hat{h}_t$.
- **RAA w/o V-self-att** denotes the variant of RAA-Net with only the 2nd-stage cross-visual attention. Without taking an early look at intrinsic cues in each visual modality with self-attention in respective V_g or V_l, RAA w/o V-self-att directly implements the mutual visual correlation.
- **RAA w/o T-V-self-att** denotes the variant of RAA-Net without "T-att" and "V-self-att" modules.
- **RAA w/o V-cross-att** denotes the variant of RAA-Net with only the 1st-stage self-visual attention. This means that the model removes the cross-visual correlation.

As shown in Table III, **RAA w/o** V_l has the largest performance degradation, which verifies the contribution of V_l. Compared with **RAA w/o** V_l, **RAA w/o** V_g improves MRR from 0.6316 to 0.6574. For this visual dialog task, local visual features (bottom-up features [21]) perform better than global visual features (VGG features). It could be that bottom-up (object-based) features can provide more fine-grained local spatial cues for visual reference. With respect to the two-stage visual-reference, the performance of **RAA w/o V-self-att** on metric Mean is around 3% lower than **RAA-Net**. The self-attention in each visual data space (V_g or V_l) comprises more comprehensive and detailed visual semantics for inferring exact answer. Compared with **RAA-Net**, **RAA w/o V-cross-att** considerably decreases MRR from 0.6683 to 0.6404, which reflects the necessity of cross-attention in our approach. Visualization examples in Figs. 3∼5 validate the effectiveness of **RAA-Net** too.

For textual reference, compared with **RAA w/o T-att**, the R@1 value of **RAA w/o T-att** decreases from 53.80 to 52.92 and the Mean value increases from 3.89 to 4.02. Reviewing history indeed benefits to discover latent and missing textual hints. Moreover, we test **RAA w/o T-V-self-att**, and its performance considerably decreases compared with **RAA w/o T-att** or **V-self-att**. It indicates that the single absence of "T-att" or "V-self-att" does not lead to a drastic performance drop, but the joint absence does.

Furthermore, there are some interesting observations. (1) The top-3 worst performances shown in Table I refer to

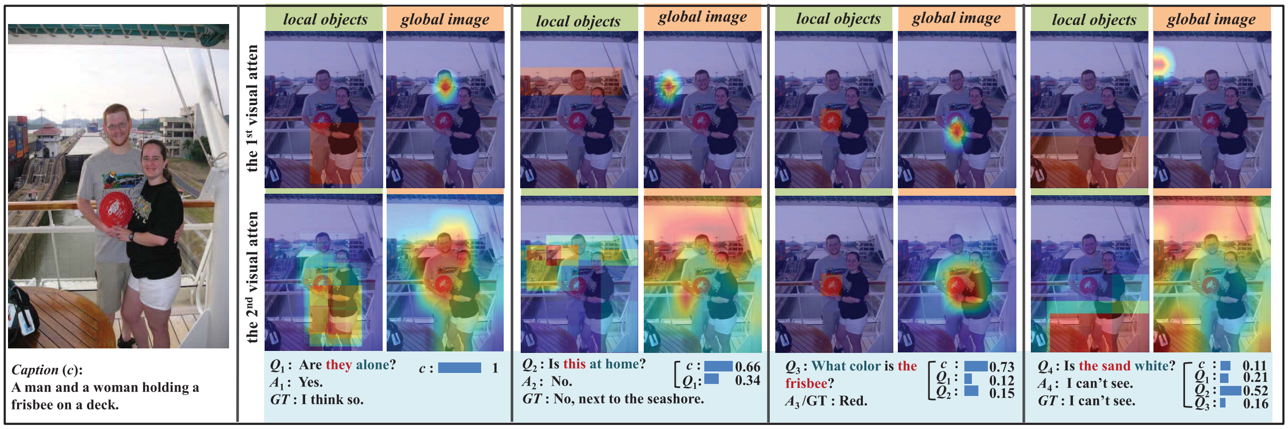

Fig. 3. Visualization of textual-visual reference-aware attention in a progressive multi-round dialog. The top layer (attention map) reflects each self-visual reference under local vision V_l and global vision V_g, while the middle-layer (attention map) describes the cross-visual reference between V_l and V_g. The bottom blue layer (*i.e.*, attention histogram) denotes the average textual attention weights of $\{h = 4\}$ multi-head attentions over the history. The following abbreviations are used: question (Q), generated answer (A), caption (c), and the ground-truth (GT).

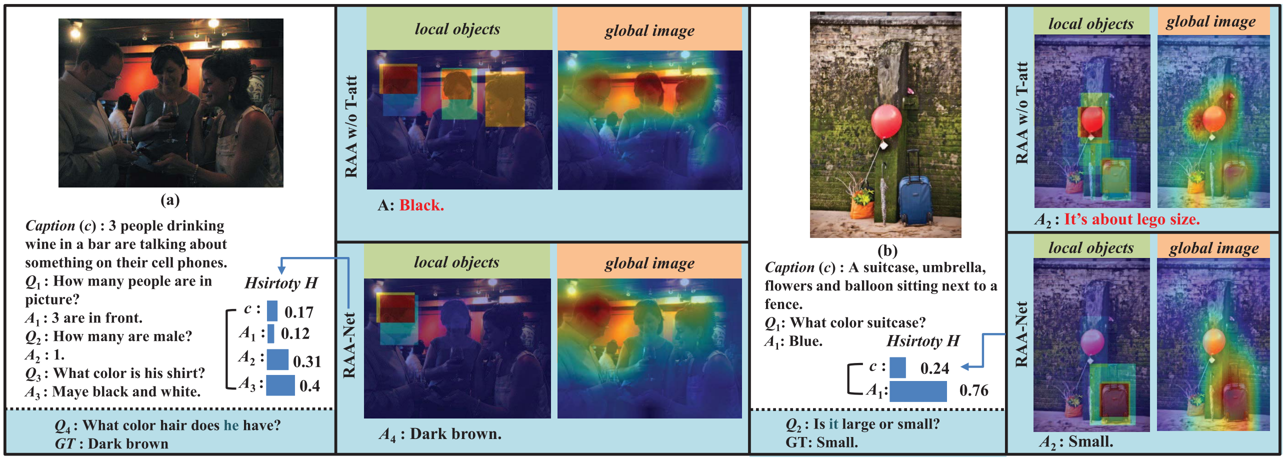

Fig. 4. Qualitative results of textual reference. Without the textual reference between Q and H, the agent prefers to attend to long sentences, *e.g.*, the caption that seems to contain more semantics, which makes the referential entities of the pronoun ("he" and "it") in the questions to be inaccurate. The weighting histogram denotes the average textual attention weights of $\{h\}$ multi-head attention over the history. This result indicates that RAA-Net can modify wrong attentive textual semantics by the co-reference between Q and H.

RAA w/o V_l, **RAA w/o V-cross-att**, and **RAA w/o V_g**. All of them have a commonality, *i.e.*, the cross-visual attention module is absent. This result confirms the importance of cross-visual reference again. (2) **RAA w/o V-cross-att** performs worse than **RAA w/o V_g**. It indicates that although V_g is a complement of V_l, if without correlation (cross-attention) between V_l and V_g, V_g may introduce noise and dominate visual reference. (3) As shown in Fig. 2, the 1st visual attention (the "V-self-att" module) is performed subsequent to the textual attention (the "T-att" module). If "T-att" is removed, the effect of "V-self-att" would be weakened. It is verified in Table III, the negative influence of **RAA w/o T-att** is more than that of **w/o V-self-att**.

C. Comparison With the State-of-the-Art Methods

1) Baselines: We compare RAA-Net with the state-of-the-art methods in both generative and discriminative settings. The differences between RAA-Net and other methods are discussed as follows. In [24], three baseline methods **LF**, **HRE**, and **MN** are introduced. Among them, **LF** directly fuses multimodal features to a joint representation to decode the answer. **HRE** uses hierarchical recurrent LSTMs [52] to separately encode the question, image, and history; finally, it adopts another LSTM to capture the temporal correlation in the entire dialog. **MN** mainly designs a memory bank to store previous dialog history. **HCIAE** [27] first attends to dialog history and then uses the attended textual features as the guidance to attend to the image. **CoAtt** [28] applies a sequential pairwise attention correlation learning. **AMEM** [32] uses an attention memory network to model the relationship between the question and dialog history. **CorefNMN** [33] relies on a weak supervision parser [53] for visual reasoning. Recently, graph-based methods have been proposed. **GNN** [30] constructs a dialog graph, where nodes are dialog entities and edges are semantic dependencies between each two nodes; the answer is regard as unobserved node that can be inferred by the EM algorithm. **FGA** [31] constructs the graph overall the

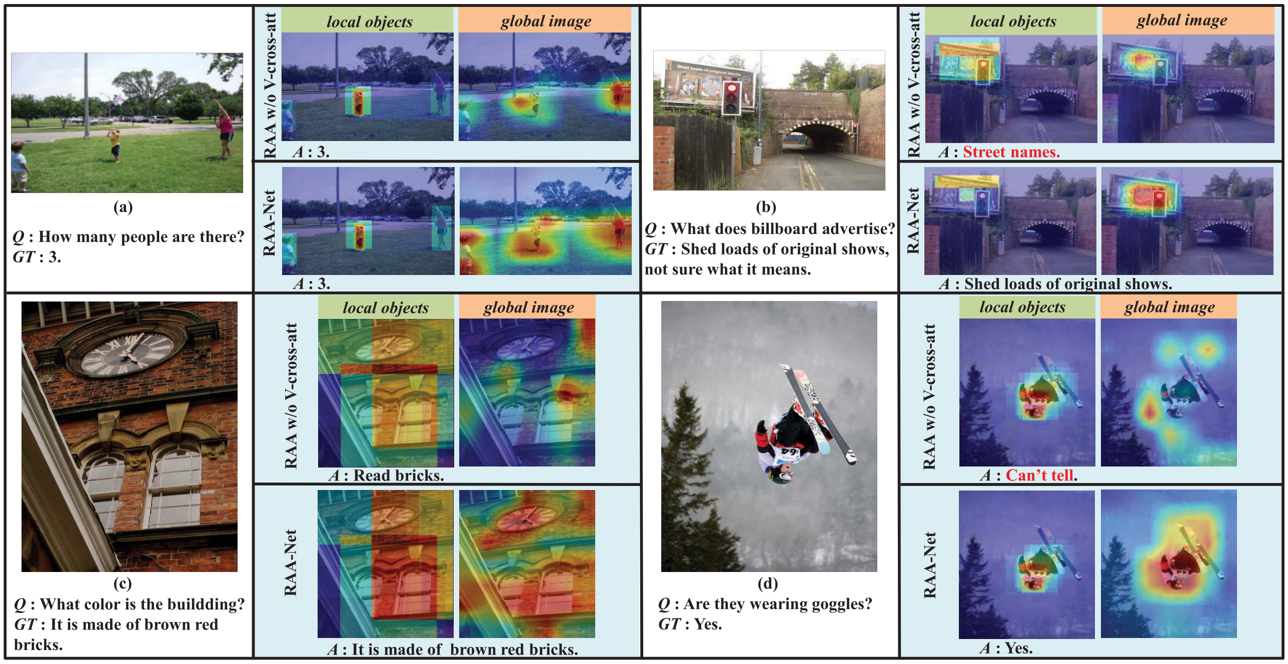

Fig. 5. Qualitative results of cross-visual attention. With the cross-visual reference, **RAA-Net** effectively discovers more latent or missing cues. These cues help the agent to infer more reasonable answers. Besides, RAA-Net prefers longer answers owing to the rich visual semantics, which includes both related objects and their surrounding spatial context.

TABLE IV
RETRIEVAL PERFORMANCE EVALUATION ON VISDIAL VAL V0.9

Discriminative Models					
	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
LF [23]	0.5807	43.82	74.68	84.07	5.78
HRE [23]	0.5846	44.67	74.50	84.22	5.72
MN [23]	0.5965	45.55	76.22	85.37	5.46
HCIAE [24]	0.6222	48.48	78.75	87.59	4.81
AMEM [29]	0.6227	48.53	78.66	87.43	4.86
GNN [27]	0.6285	48.95	79.65	88.36	4.57
CoAtt [25]	0.6398	50.29	80.71	88.81	4.47
CorefNMN [30]	0.6410	50.92	80.18	88.81	4.45
FGA [28]	0.6525	51.43	82.08	89.56	4.35
RAA-Net	**0.6683**	**53.80**	**82.99**	**90.86**	**3.89**
Generative Models					
	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
LF [23]	0.5199	41.83	61.78	67.59	17.07
HRE [23]	0.5237	42.29	62.18	67.92	17.07
MN [23]	0.5259	42.29	62.85	68.88	17.06
HCIAE [24]	0.5386	44.06	63.55	69.24	16.01
CoAtt [25]	0.5411	44.32	63.82	69.75	16.47
RAA-Net	**0.5593**	**46.61**	**65.40**	**71.15**	**14.87**

multi-modal features and estimates their interactions by factor graph attention.

2) Results on VisDial v0.9: The discriminative comparison is shown in Table IV; RAA-Net performs much better than other methods. Except for the proposed RAA-Net, CorefNMN, CoAtt, GNN, and FGA are the top-4 best performing methods. **CorefNMN** [33] tackles only the visual reference in global visions. By considering both local and global visions, **RAA-Net** considerably outperforms **CorefNMN** and results in an increase in MRR from 0.6410 to 0.6683. In **CoAtt** [28], a complicated co-attention is implemented by multiple mutual interactions among the question, history, and global visual features. **RAA-Net** just progressively imposes question semantics on history and visual features. Compared with the R@1 value of **RAA-Net** at 53.80, **CoAtt** decreases to 50.29. This result indicates that our progressive textual-visual attention is much more effective than the complicated mutual attention interaction between every two modalities in **CoAtt**.

Next, compared with **GNN** [30], our model improves the MRR metric by approximately 4%. **GNN** explores the graph-based textual reference, while **RAA-Net** handles mutual textual-visual reference learning. FGA [31] is the state-of-the-art graph-based method for visual dialog, which treats the candidate answer embedding feature A_t as new context cue and introduces it into the multi-modal encoding training. Without utilizing these candidate answers in the training process, our model still produces better results, *e.g.*, decreases Mean from 4.35 to 3.89. Regarding the generative model, **RAA-Net** also achieves the best performance.

3) Results on VisDial v1.0: We also evaluate **RAA-Net** on the VisDial v1.0 dataset. The comparison results are shown in Table V. Except for **FGA** [31], **RAA-Net** has better performance than other methods. Comparable with **FGA**, our model performs better in terms of the Mean and NDCG. It is worth noting that in VisDial v1.0, there is a new metric, *i.e.*, NDCG. NDCG is a widely adopted significant indicator, which involves comprehensive quantitative semantics evaluation. Other metrics are mainly influenced by the rank of the correct answer in the candidate answer list, while NDCG measures the semantic similarity of the output answer list. The NDCG score will be high if more semantically relevant answers appear in the top positions in the answer list. As shown in Table V, **RAA-Net** achieves the best NDCG performance. For example, compared with the latest methods

TABLE V

RETRIEVAL PERFORMANCE EVALUATION OF DISCRIMINATIVE MODELS ON VISDIAL TEST-STD V1.0

Model	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓	**NDCG↑**
LF [23]	0.5542	40.95	72.45	82.83	5.95	45.31
HRE [23]	0.5416	39.93	70.45	81.50	6.41	45.46
MN [23]	0.5549	40.98	72.30	83.30	5.92	47.50
MN-att [23]	0.5690	42.43	74.00	84.35	5.59	49.58
LF-att [23]	0.5707	42.08	74.83	85.05	5.41	49.76
GNN [27]	0.6137	47.33	77.98	87.83	4.57	52.82
FGA [28]	**0.6370**	**49.58**	**80.97**	88.55	4.51	52.10
CorefNMN [30]	0.6150	47.55	78.10	88.80	4.40	54.70
RAA-Net	0.6286	49.05	79.65	**88.85**	**4.35**	**55.42**

GNN and **FGA**, the NDCG performance of our model is improved by 2.6% and 3.3%, respectively.

D. Qualitative Results

To further demonstrate the interpretability of our solution, we provide an example in Fig. 3, which progressively describes the semantic augmentation by both textual and visual referring in multi-round QA pairs. For the question Q_1 "Are they alone?", in the 1st self-visual reference stage, global visual attention is located in a small relevant region (the head of "a man"), while local attention focuses on some areas that partially cover "a man" and "a woman." Then, in the 2nd cross-visual reference, both local and global attention maps are modified to highlight the areas covering the body of the two people. The visualization results indicate the effectiveness of RAA-Net. For questions Q_2 and Q_3, visual cues have the right consistency in both local and global visions. Visual consistency enhances the justification of the semantic reasoning process. However, in some cases, local and global sights focus on different regions. Because the object ("sand") in question Q_4 does not appear in the image, both local and global attention maps look over different regions and try to provide a more comprehensive scene parsing. The complementarity of local and global visions helps the agent to infer a correct answer.

In addition, we discuss the influence of the textual reference in the proposed model. The attention maps in Fig. 4 are obtained with and without the multi-head textual attention module, corresponding to **RAA-Net** and **RAA w/o T-att**. There are two examples in Fig. 4. It is easy to find that without the textual reference between Q and H, the referential entities of the pronoun ("he" and "it") in the questions are inaccurate. As shown in Fig. 4 (a), QA pairs in history always talk about all three people. Without textual inferring, the agent provides an implicit consent to the same objects (three people) and infers a wrong answer (*i.e.*, "black") to question "Q_4: What color hair does he have?". Fig. 4 (b) shows more obvious attention differences. **RAA w/o T-att** focuses on all subjects in the caption, while **RAA-Net** mainly attends to the subject "suitcase" related to question "Q_2: Is it large or small?". Without a textual reference using multi-head textual attention, the model prefers to attend to long sentences, *e.g.*, captions that seem to contain more semantics. However, long sentences may introduce useless textual semantics or semantic noise. Fig. 4 shows the visualization of the average textual attention histogram under multi-heads. The textual attention distribution indicates that **RAA-Net** modifies the wrong attentive textual semantics indicated by the co-reference between Q and H. In summary, the textual reference-aware process is necessary.

Finally, because the relationship within a single visual feature sequence (intra-relationship) has been already discussed in many studies, here, we focus on the discussion of the influence of inter-attention (cross-attention) among different feature sequences. As shown in Fig. 5, there are four examples. In Fig. 5 (a) and (b), the local and global attention maps are consistent in the relevant regions. For inferring a correct answer, **RAA w/o V-cross-att** still concentrates on some specific objects, whereas **RAA-Net** expands to consider the more surrounding spatial context. With respect to the different attention responses from local and global views in Fig. 5 (c) and (d), **RAA w/o V-cross-att** introduces all these original visual cues into the final multimodal fusion, while **RAA-Net** remedies their mutual visual complementarity by cross-visual correlation. Experimental results confirm the effectiveness of the cross-correlation process in **RAA-Net**. We can see that after the cross-correlation operation, both local and global responses become consistent, which shows the correct answer more explicitly. In addition, compared with **RAA w/o V-cross-att**, **RAA-Net** tends to generate longer answers owing to the rich visual semantics, involving both related objects and their surrounding spatial context.

V. CONCLUSION

In this study, we propose a textual-visual Reference-Aware Attention Network (RAA-Net) for performing visual dialog task. It provides a fine-grained understanding of the multi-modality context. In RAA-Net, we realize the textual-visual reference through a multi-head textual attention mechanism and a two-stage visual reference involving both self- and cross-visual correlation learning. Experimental results on VisDial v0.9 and v1.0 datasets demonstrate that the proposed model achieves state-of-the-art performance and shows explainable visualization results.

REFERENCES

[1] Y. Hua, S. Wang, S. Liu, A. Cai, and Q. Huang, "Cross-modal correlation learning by adaptive hierarchical semantic aggregation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1201–1216, Jun. 2016.

[2] F. Zhang, F. M. Moss, R. Baddeley, and D. R. Bull, "BVI-HD: A video quality database for HEVC compressed and texture synthesized content," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2620–2630, Oct. 2018.

[3] S. Yang, L. Li, S. Wang, W. Zhang, Q. Huang, and Q. Tian, "SkeletonNet: A hybrid network with a skeleton-embedding process for multi-view image representation learning," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2916–2929, Nov. 2019.

[4] S. Wang, Y. Chen, J. Zhuo, Q. Huang, and Q. Tian, "Joint global and co-attentive representation learning for image-sentence retrieval," in *Proc. ACM Multimedia Conf. (MM)*, 2018, pp. 1398–1406.

[5] G. Song, S. Wang, Q. Huang, and Q. Tian, "Harmonized multi-modal learning with Gaussian process latent variable models," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 17, 2019, doi: 10.1109/TPAMI.2019.2942028.

[6] A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.

[7] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.

[8] T. Chen *et al.*, "'Factual' or 'emotional': Stylized image captioning with adaptive learning and attention," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 527–543.

[9] M. Yang *et al.*, "Multitask learning for cross-domain image captioning," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1047–1061, Apr. 2019.

[10] N. Yu, X. Hu, B. Song, J. Yang, and J. Zhang, "Topic-oriented image captioning based on order-embedding," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2743–2754, Jun. 2019.

[11] S. Ye, J. Han, and N. Liu, "Attentive linear transformation for image captioning," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5514–5524, Nov. 2018.

[12] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell, "Natural language object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4555–4564.

[13] 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.*, 2016, pp. 457–468.

[14] H. Zhang, Y. Niu, and S.-F. Chang, "Grounding referring expressions in images by variational context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4158–4166.

[15] C. Deng, Q. Wu, Q. Wu, F. Hu, F. Lyu, and M. Tan, "Visual grounding via accumulated attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7746–7755.

[16] Z. Yu, J. Yu, C. Xiang, Z. Zhao, Q. Tian, and D. Tao, "Rethinking diversified and discriminative proposal generation for visual grounding," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1114–1120.

[17] D. Liu, H. Zhang, Z.-J. Zha, and F. Wu, "Learning to assemble neural module tree networks for visual grounding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4672–4681.

[18] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[19] M. Ren, R. Kiros, and R. S. Zemel, "Exploring models and data for image question answering," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 2953–2961.

[20] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded question answering in images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4995–5004.

[21] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[22] B. Patro and V. P. Namboodiri, "Differential attention for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7680–7688.

[23] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville, "GuessWhat?! Visual object discovery through multi-modal dialogue," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5503–5512.

[24] A. Das *et al.*, "Visual dialog," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 326–335.

[25] D. Guo, H. Wang, and M. Wang, "Dual visual attention network for visual dialog," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4989–4995.

[26] D. Guo, H. Wang, H. Zhang, Z. Zha, and M. Wang, "Iterative context-aware graph inference for visual dialog," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020.

[27] J. Lu, A. Kannan, J. Yang, D. Parikh, and D. Batra, "Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 314–324.

[28] Q. Wu, P. Wang, C. Shen, I. Reid, and A. van den Hengel, "Are you talking to me? Reasoned visual dialog generation through adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6106–6115.

[29] T. Yang, Z.-J. Zha, and H. Zhang, "Making history matter: History-advantage sequence training for visual dialog," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2561–2569.

[30] Z. Zheng, W. Wang, S. Qi, and S.-C. Zhu, "Reasoning visual dialogs with structural and partial observations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6669–6678.

[31] I. Schwartz, S. Yu, T. Hazan, and A. G. Schwing, "Factor graph attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2039–2048.

[32] P. H. Seo, A. Lehrmann, B. Han, and L. Sigal, "Visual reference resolution using attention memory for visual dialog," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 3719–3729.

[33] S. Kottur, J. M. F. Moura, D. Parikh, D. Batra, and M. Rohrbach, "Visual coreference resolution in visual dialog using neural module networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 153–169.

[34] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 39–48.

[35] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[36] K. Fang, A. Toshev, L. Fei-Fei, and S. Savarese, "Scene memory transformer for embodied agents in long-horizon tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 538–547.

[37] K.-M. Kim, S.-H. Choi, J.-H. Kim, and B.-T. Zhang, "Multimodal dual attention memory for video story question answering," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 673–688.

[38] M. R. Farazi and S. Khan, "Reciprocal attention fusion for visual question answering," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 1–12.

[39] P. Lu, H. Li, W. Zhang, J. Wang, and X. Wang, "Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7218–7225.

[40] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8927–8936.

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

[42] L. J. Ba, R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*. [Online]. Available: https://arxiv.org/abs/1607.06450

[43] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 91–99.

[44] R. Krishna *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

[46] D. Teney, P. Anderson, X. He, and A. van den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4223–4232.

[47] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[48] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[50] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[51] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, 2019, pp. 4171–4186.

[52] I. V. Serban *et al.*, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 3295–3301.

[53] 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. Vis. (ICCV)*, Oct. 2017, pp. 804–813.

Dan Guo received the B.E. degree in computer science and technology from Yangtze University, China, in 2004, and the Ph.D. degree in system analysis and integration from the Huazhong University of Science and Technology, China, in 2010. She is currently an Associate Professor with the School of Computer Science and Information Engineering, Hefei University of Technology, China. Her research interests include computer vision, machine learning, and intelligent multimedia content analysis.

Hui Wang received the B.E. degree in computer science and technology from the Hefei University of Technology, China, in 2018, where he is currently pursuing the M.S. degree in computer technology. His current research interests include computer vision and deep learning.

Shuhui Wang received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2006, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in China, 2012. He is currently an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include vision-language, visual learning, and multimedia retrieval.

Meng Wang (Senior Member, IEEE) received the B.E. and Ph.D. degrees in the special class for the gifted young from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the Hefei University of Technology, China. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He has authored more than 200 book chapters, journal, and conference papers in his research areas. He was a recipient of the ACM SIGMM Rising Star Award 2014. He is an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.