Depth-aware and Semantic Guided Relational Attention Network for Visual Question Answering

Yuhang Liu, Wei Wei†, Daowan Peng, Xian-Ling Mao, Zhiyong He, Pan Zhou

***Abstract*—Visual relationship understanding plays an indispensable role in grounded language tasks like visual question answering (VQA), which often requires precisely reasoning about relations among objects depicted in the given question. However, prior works generally suffer from the deficiencies as follows, (1) spatial-relation inference ambiguity, it is challenging to accurately estimate the distance of a pair of visual objects in 2D space if there is a visual-overlap between their 2D bounding-boxes, and (2) language-visual relational alignment missing, it is insufficient to generate a high-quality answer to the question if there is a lack of alignment in the language-visual relations of objects during fusion, even using a powerful fusion model like Transformer. To this end, we first model the spatial relation of a pair of objects in 3D space by augmenting the original 2D bounding-box with 1D depth information, and then propose a novel model named Depth-aware Semantic Guided Relational Attention Network (DSGANet), to explicitly exploit the formed 3D spatial relations of objects in an intra-/inter-modality manner for precise relational alignment. Extensive experiments conducted on the benchmarks (VQA v2.0 and GQA) demonstrate DSGANet achieves competitive performance compared to pretrained and non-pretrained models, such as 72.7% vs. 74.6% based on the learned grid features on VQA v2.0.**

***Index Terms*—Visual question answering, relational reasoning, depth estimation, multi-modal representation.**

I. Introduction

Recently, cross-modal problem has received a considerable amount of attentions from both computer vision (CV) and natural language processing (NLP) communities, which requires to simultaneously span both modalities (*i.e.,* vision and language) for achieving domain-specific tasks, such as visual question answering (VQA) [1]–[6], the goal of which is to answer questions about images through fully understanding of the semantics of the input text-image, and generating the correct answer to the question.

Generally, VQA methods attempt to answer the question according to the visual clues mined from the image and the semantics of the corresponding question. Prior studies have explored attention mechanisms, such as co-attention based [7], densely-connected attention-based [8], and modular co-attention based [9]. Another line of works have explored to learn cross-modal alignment via pretraining with large-scale unlabeled data [10]–[14]. Despite effectiveness, there is a natural deficiency of visual-reasoning ability for such methods to answer the questions requiring in-depth understanding of the spatial relations among visual objects. To address this, there exist several attempts on neural module network to decompose the input question into several self-designed functions to answer the questions step-by-step, such as heterogeneous modules [15], [16] and homogeneous modules [17]–[19]. Nevertheless, these methods still easily fail to achieve the VQA task due to heavily relying on the handcrafted modules [20]. Instead, several approaches utilize the monolithic networks to enrich the visual features with the multimodal/contextual information required for reasoning, such as graph neural network [21]–[24] and Transformer [25] which has become an effective and widely-used solution as its well-designed modules (e.g., self-attention) achieve the promising performance on cross-model alignment (inter-model) and contextual information acquisition (intra-modal). Following the success of Transformer, many variants of Transformer-based VQA models are proposed [2], [9], [26].

Nevertheless, the vanilla Transformer with self-attention module is far from enough for visual reasoning during inference (especially for spatial-relation related questions), since it solely relies on visual features to measure correlations of two objects. Indeed, several works have already attempted to make use of 2D bounding-box relations [21], [26], [27] for learning the representation of spatial-relations of objects

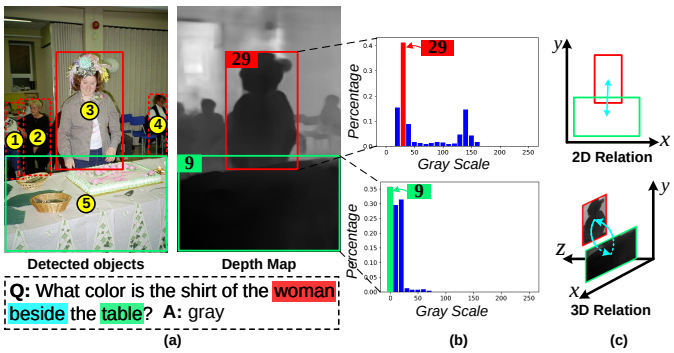

Fig. 1. (a) An example of image-question-answer triplet and its corresponding pixel-wise depth map (brighter means farther in distance). (b) The grayscale histograms of corresponding objects. The number indicates the peak gray value of each region. (c) An illustration of 2D&3D spatial relations. x, y, z-axies denote the directions along the width, height and depth of the image.

Wei Wei is the corresponding author.

Yuhang Liu, Wei Wei and Daowan Peng was with Cognitive Computing and Intelligent Information Processing (CCIIP) Laboratory, School of Computer Science and Technology,Huazhong University of Science and Technology, Wuhan, China. E-mail: {lyuhang,weiw,pengdw}@hust.edu.cn.

Xian-Ling Mao was with School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. E-mail: maoxl@bit.edu.cn.

Zhiyong He was with Naval University of Engineering, Wuhan, China. E-mail: moonmon_pub@outlook.com.

Pan Zhou was with Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology, Wuhan, 430074, Hubei, China. E-mail: panzhou@hust.edu.cn.

by narrowing down the gap between their visual features and linguistic semantics. However, they may even deteriorate the performance. Without loss of generality, we take the example in Figure 1-(a) for illustration, in Figure 1-(a), for answering the question, VQA models apparently need to infer the relative distance of the table to the four different women for distinguishing. However, previous studies modeling spatial-relations in 2D space (Figure 1-(c) (up)) may mislead the alignment of the spatial relation ($r^{bbox}_{<woman,table>}$) and the linguistic semantics ($r^{semantic}_{beside}$).

Inspired by the observation, in this paper we consider to model the spatial relation of two objects in 3D space instead of 2D bounding-box via augmenting with 1D depth information (as shown in Figure 1-(c) (down)), for accurately inferring the relative relations of different visual objects. Therefore, the distances between the table to the four different women can be calculated more accurately by means of depth information, and thus we can find that woman-3 is more closely to the table as compared to the others (*i.e., woman*-1,2,4). Additionally, we also propose an innovative relation alignment model, named Depth-aware and Semantic Guided Relational Attention Network (DSGANet), to accurately locate the target visual objects through fully exploiting the formed 3D visual relations for relation alignment during visual-language features fusion. In contrast to self-attention that derives correlations only based on feature similarities, DSGANet is capable of capturing the spatial relevant context for accurate relation reasoning, via estimating the distance of a pair of objects in 3D space. Furthermore, we claim that our proposed DSGANet can be applicable to two different types of visual features, *i.e.,* objects [1] and grids [2], and achieve competitive performance in terms of different evaluation metrics.

In summary, the main contributions are the following,

- To the best of our knowledge, this is the first attempt that explicitly builds 3-dimensional spatial-relations between objects, and performs cross-modal relational alignments for more accurate visual reasoning about the objects depicted in the given question.
- We propose a Depth-aware and Semantic Guided Relational Attention Network (DSGANet) to precisely capture the spatial context via modeling relational alignment in an inter-/inter-modality manner simultaneously. Meanwhile, we also evaluate the effectiveness of joint vision and language understanding by our proposed model equipped with two different types of grid-based features, and which work surprisingly well.
- We conduct extensive experiments to evaluate the effectiveness of our proposed DSGANet, in which our proposed model achieves competitive performance compared to pretrained and non-pretrained models over different datasets, for example, 74.06% overall accuracy on VQA v2.0 and 58.32% on GQA.

II. Related Work

A. Visual Question Answering

Generally, VQA aims at answering a question to its corresponding given image. Indeed, there already exist several early works in VQA domain to research on fusion strategies [4]–[6], [28] and attention mechanisms [1], [29] to preserve fine-gained feature of images for joint visual-language representation learning, *e.g.,* small objects [29] or question-relevant regions [1]. Despite their significant improvement, it remains challenging to answer the questions that involve multiple objects and require visual reasoning for grounding. To solve this problem, there already exist several efforts dedicated to research on two parts: task-decomposing and graph-based reasoning. Neural module networks (NMN) [15], [17]–[19] decompose questions into sub-tasks and accordingly compose neural modules to sequentially answer the questions. These works perform well on synthetic datasets and preserve the compositional interpretability. However, they are not widely adopted in real-world datasets, which are characterized by open vocabulary and require more reasoning abilities. In contrast to NMN that explicitly decomposes questions, graph-based methods perform implicit visual reasoning along the scene graph of images via message passing [21]–[24]. The resulting regional features are fully contextualized and distinguishable for object groundings. In our work, we follow the graph-based framework with Transformer [9], and build graph networks with explicit 3D spatial relationships to learn contextual object representations in 3D space.

B. Visual Relationship Modeling

Visual relationship modeling plays an important role in image understanding. Currently, there are tremendous related works focusing on the task of visual relation extraction [30], which brings a great impact on a variety of visual comprehension tasks, such as change captioning [31], visual question answering [32], *etc.* Indeed, relationships encode the interactions among objects, which are vital in locating the targets via contextual information. However, existing VQA methods simplify this process with *implicit relations* [21], [24], [33] and *explicit relations* [21], [22], [27], [34]. *Implicit relations* are derived from the feature correlations (*e.g.,* self-attention) and *explicit relations* denote as the geometric or semantic relationships between objects. Limited by computational efficiency, *implicit relations* are widely adopted in the state-of-the-art methods [9], [21], [24]. However, the ignorance of input structure limits their ability for visual reasoning. In contrast, *explicit relations* take account of geometric or semantic relationships between objects. For example, a few works [21], [26], [27], [32], [35] attempt to build geometric relations with bounding boxes and achieve significant performance. However, it's still intractable for 2D bounding-boxes to represent the real world spatial relations (*e.g.,* the distance in 3D space). In this work, we extend 2D bounding-box relations between objects with 1D *depth* information and learn the relational alignment between vision and language with graph-based framework.

C. Depth Estimation

Depth estimation has been a popular research area due to its importance in the understanding of 3D world [36], [37].

For example, the depth information shares common knowledge with semantic segmentation which can be transferred to semantic segmentation task [38] and boost the performance. In addition, 2D images lose the depth information, which could lead to ambiguity when inferring positions of the objects in the visual scenes. Scene Graph based Change Captioning (SGCC) proposed to describe the relative position relationship via 3D information of images, which overcomes the disturbances from viewpoint changes [31]. Depth-aware MGAT devised depth information to distinguish different objects, improving the answering of counting-related questions [39]. The recently proposed work [40] exploited similar depth information to our method. The main difference is that [40] regards the depth information as weakly supervised labels which are used for pretraining the models. However, there is no corresponding module designed for modeling the 3-dimensional relationships between objects, which limits the visual reasoning capabilities of the model. To address this issue, we propose a depth-aware and semantic guided relational attention mechanism, which explicitly models the 3-dimensional spatial relations and learns the cross-modal relational alignment with graph-based framework.

D. Visual Features

It has been shown that recent advances in VQA benefit to a certain extent from the visual representations of images [41]. The visual features used in VQA can be divided into three categories, *i.e., global*, *objects* and *grids*. *Global* method directly encodes the image into a global feature vector [28], [42], [43], which fails to attain fine-grained information about the image. Therefore, *objects* have been the de facto choice for most VQA models [1], [3], [9], [11], [21], [26], [29], [44]. Specifically, a pre-trained Faster R-CNN [45] is used to detect objects from the images, resulting in a set of object features and bounding boxes. In common practice, the object detection model is pre-trained in advance on Visual Genome dataset [46] with annotations of object classes and attributes. The objects act as visual priors and promote VQA models to focus on salient regions. Due to the pre-training strategy, the object features contains semantics of object categories and attributes, which facilitates cross-modal alignment. Since object-based features [1] were proposed, they have been widely used in the subsequent researches [3], [6], [21], [44]. However, Jiang et al. [2] observed that the key factor contributing to the good performance with bottom-up features [1] do not rely on the feature format (*i.e., object* or *grid*). Therefore, they proposed *grid* features, which skips the expensive region-related steps and directly uses C_4 output of Faster R-CNN backbone as visual input. Their experiments show that the grid features can perform competitive and even better than object features with less inference time.

Nevertheless, they simply replace the object input with grid features, which still ignores the spatial relations between *grids*. To demonstrate the generalization of our proposed attention mechanism, we adapt our DSGA to *grid* features, which brings a significant improvement.

III. Preliminary

In this section, we first give the statement of our visual question answering problem (Section III-A), and then present an overview of our proposed model (Section III-B). For clarity, some notations and their definitions are listed in Table I.

A. Problem Statement

Let I, Q and $\mathcal{A} = \{a_i\}_{|\mathcal{A}|}$ be the image, the question grounded on I and the candidate answer set respectively, where a_i is an answer in $\mathcal{A}$; $\mathcal{O} = \{o_i\}_{|\mathcal{O}|}$ denotes the visual objects extracted from I; and r_{ij} indicates the spatial relation between object o_i and object o_j. Hence, the problem of visual question answering aims at selecting a correct answer from candidate answer set $\mathcal{A}$ when given a text-image input $< I, Q >$, which can be formed as a classification problem as follows,

$$a^* \leftarrow \arg\max_{a \in \mathcal{A}} \Pr(a|I, Q) \tag{1}$$

B. Overview

Next, we present an overview (as shown in Figure 2) of our approach to addressing the visual question answering problem. Specifically, it consists of 4 components, namely, (i) **Image Representation**, (ii) **Question Representation**, (iii) **Encoding Module**, (iv) **Prediction Module**.

Figure 2 gives a brief illustration about our proposed DSGANet: (1) *Image Representation* extracts regional features and depth map from image I, which are used to construct and initialize a fully-connected object graph $\mathcal{G}$ with region features $\mathcal{V}$, implicit relations $\mathcal{E}^{imp}$, and explicit spatial relations $\mathcal{E}^{spa}$. (2) *Question Representation* encodes questions with Bert [47] or GRU, resulting in a set of word embeddings $\mathbf{S} = \{s_i\}_{i=0}^{M-1}$, where M denotes the question length. (3) *Encoding Module* adopts Transformer framework to refine region features with context-aware information under the guidance of question semantics, in which Depth-aware and Semantic Guided Attention (DSGA) mechanism is proposed to take 3D spatial relations between objects into account, facilitating the visual

TABLE I
NOTATIONS AND DEFINITIONS

Notation	Definition
Input and Output	
I	The input image.
Q	The input question.
$\mathcal{A}$	The candidate answer set, $\{a_i\}_{\|\mathcal{A}\|}$.
Object Graph	
$\mathcal{G}$	The fully-connected object graph, $\mathcal{G} = \{\mathcal{V}, \mathcal{E}^{imp}, \mathcal{E}^{spa}\}$.
$\mathcal{V}$	The object set of graph $\mathcal{G}$.
$\mathcal{E}^{imp}$	The implicit relation set of graph $\mathcal{G}$.
$\mathcal{E}^{spa}$	The explicit spatial relation set of graph $\mathcal{G}$.
$\mathbf{b}$	The bounding-box features.
$\mathbf{dep}_i^{hist}$	The histogram vector of i-th object.
dep_i	The grayscale with the highest frequency.
$\mathbf{r}_{i,j}^{bbox}$	The *bbox* relation of i, j-th objects.
$\mathbf{r}_{i,j}^{depth}$	The *depth* relation of i, j-th objects.

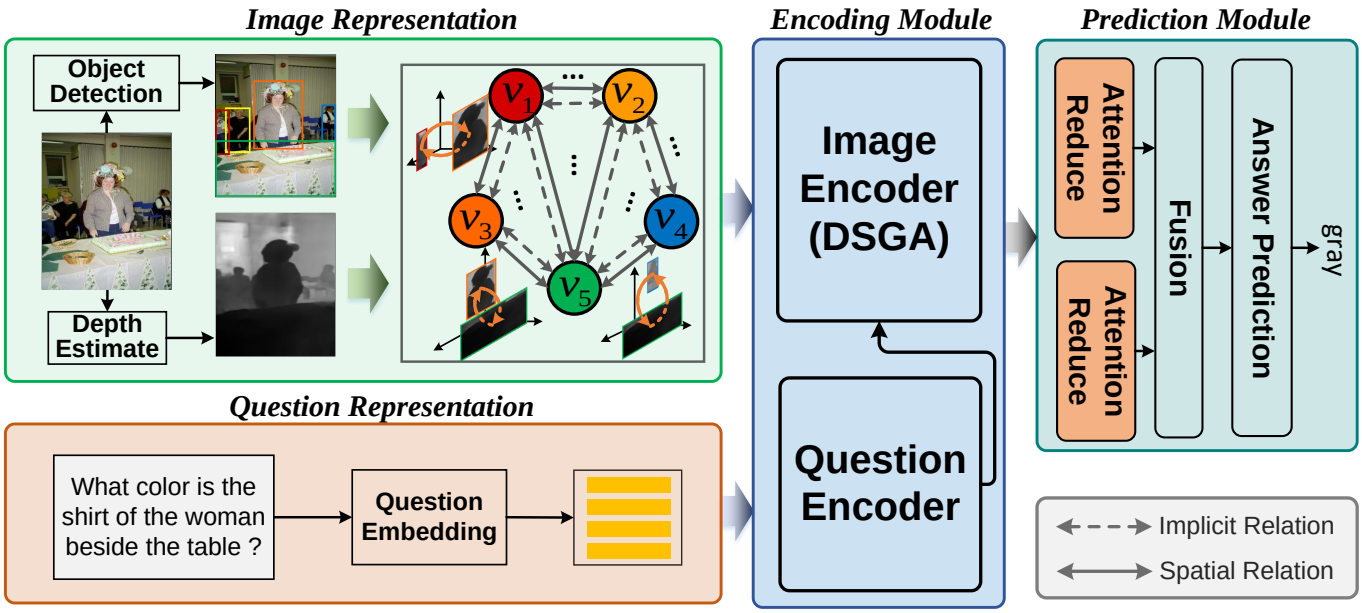

Fig. 2. The overall architecture of our proposed model, which consists of 4 modules, *i.e., Image Representation*, *Question Representation*, *Encoding Module* and *Prediction Module*.

reasoning about spatial relationships. (4) *Prediction Module* combines the features from both image and question, which is used to predict the final answer.

IV. Depth-aware and Semantic Guided Relational Attention Network

In the following sections, we first formulate the image as a fully-connected object graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}^{imp}, \mathcal{E}^{spa}\}$, so that relational visual reasoning can be performed among the objects (Section IV-A). Specifically, $\mathcal{G}$ represents regions as nodes and constructs implicit&explicit relations for edges. Then, the question is embedded into a set of semantic vectors using BERT or GRU (Section IV-B). Thereafter, a Transformer-based architecture is exploited to encode the question representations and the fully-connected object graph (Section IV-C). In the encoding module, we present Depth-aware and Semantic Guided Relational Attention (DSGA), which is incorporated with Transformer to perform visual reasoning via message passing on $\mathcal{G}$. Finally, the output features from image and question are combined to predict the final answer using multi-class classification (IV-D).

A. Image Representation

The image is represented as a fully-connected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}^{imp}, \mathcal{E}^{spa}\}$, in which the nodes $\mathcal{V}$ are initialized using the object features $\mathcal{O} = \{o_i\}_{|\mathcal{O}|}$ extracted from Faster R-CNN, and the edges $\mathcal{E}$ are built from our proposed implicit and explicit 3D spatial relations.

Node Initialization. The nodes $\mathcal{V}$ are initialized with *object* or *grid* features, and each node additionally contains 2D position vector $\mathbf{b}$ as well as 1D depth feature dep_i, which can be used for explicit 3D spatial relationship modeling.

Object features are generated via bottom-up attention [1], which produces N 2048-dimensional RoI features and bounding-boxes specified by coordinates $\mathbf{b} = (x^{tl}, y^{tl}, x^{br}, y^{br})$, where (x^{tl}, y^{tl}) and (x^{br}, y^{br}) are the top-left and bottom-right corners of the bounding-box. Each feature vector and its corresponding bounding-box are concatenated to initialize node features. As for depth features, we exploit depth estimation model [37] to predict the pixel-wise depth map of the image, and scale the values to the range of 0 to 255 (Fig.1-(a)). To obtain the 1D depth feature for each object, the grayscale histogram of i-th object is generated from its bounding-box region (Fig.1-(b)), and the grayscale value with the highest frequency is regarded as the object depth. However, we find that the histogram with 256 bins tends to produce multiple peaks, which could lead to inaccurate depth estimation if we simply choose the highest frequency. In addition, 256-bins histogram are redundant to represent the depth distribution since most of the grayscale values are concatenated in a few bins. As a result, we perform a "smooth" mechanism using fewer bins (*i.e.,* 16 bins) for histogram, which estimates the averaged depth in each bounding box. Specifically, the grayscale values are divided into k ($k = 16$) bins (0∼15, 16∼31, ..., 240∼255), producing a k-dimensional frequency vector $\mathbf{dep}_i^{hist} \in [0, 1]^k$. The gray value $dep_i \in [0, 255]$ with the highest frequency is regarded as the object depth (refer to Figure 1 for illustration).

Grid features are extracted from the output of ResNet C_5 layer [2], producing HW 2048-dimensional feature vectors, where H, W indicates the height and width of the feature map. Each grid is assigned with a Cartesian coordinate $\mathbf{b} = (x, y)$ (*e.g.,* $(0, 0), (0, 1), ..., (W-1, H-2), (W-1, H-1)$) for 2D position representation. To obtain the 1D depth feature for each grid, we associate each grid to a square region in the original image. Specifically, we split the depth map into

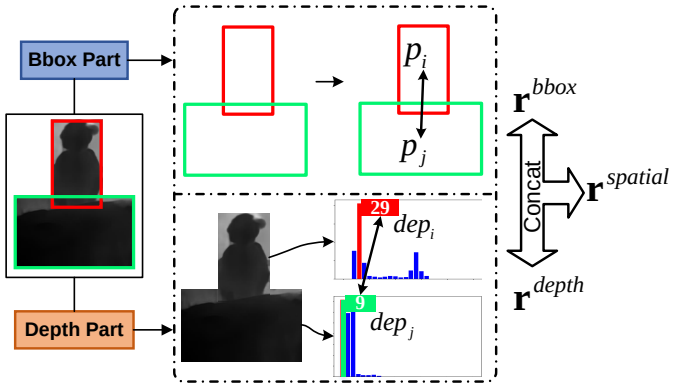

Fig. 3. Illustrations for spatial relationship modeling. The spatial relation consists of two parts: the *bbox relation* estimates the relative position and size, and the *depth relation* approximates the visual distance.

$H \times W$ regions. Similar to *object* features, the depth features $\mathbf{dep}^{hist}$ and dep for each *grid* can be generated from the corresponding region.

Implicit Relations. Formally, we treat the self-attention (SA) in the vanilla Transformer as the implicit relations $\mathcal{E}^{imp}$ for object pairs. Without loss of generality, the nodes $\mathcal{V}$ are embedded and transformed to a set of N d_h-dimensional vectors, which is packed into a matrix $\mathbf{X} \in \mathbb{R}^{N \times d_h}$. We follow MCAN [9] to implement SA on $\mathbf{X}$, which generates queries $\mathbf{Q} \in \mathbb{R}^{N \times d_h}$, keys $\mathbf{K} \in \mathbb{R}^{N \times d_h}$ and values $\mathbf{V} \in \mathbb{R}^{N \times d_h}$. Thereafter, the correlation score of x_i and x_j is calculated by scaled dot-product of the corresponding query q_i and key k_j. In matrix form, the equation is described as follows:

$$\begin{aligned} \mathrm{SA}(\mathbf{X}) &= \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_h}}\mathbf{V} \\ \mathbf{Q} &= \mathbf{W}^Q\mathbf{X} \\ \mathbf{K} &= \mathbf{W}^K\mathbf{X} \\ \mathbf{V} &= \mathbf{W}^V\mathbf{X} \end{aligned} \tag{2}$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ are fully-connected layers that project input $\mathbf{X}$ into queries, keys, and values, respectively. According to equation 2, it can be observed that correlation scores depend on feature similarities. It means that the regional features will gather contextual messages with similar semantics (*e.g., color*, *shape*, *semantics*, *etc.*), which is insufficient for image encoding due to the structural nature of images. Hence, we propose to construct explicit 3D spatial relations for object pairs, which captures the spatial dependency between objects more effectively.

Spatial Relations. We extend the 2D bounding-box with 1D depth information to construct 3D spatial relations $\mathcal{E}^{spa}$. The augmented spatial relations facilitate the comprehension of the visual distance and help to narrow the relational gap between 2-dimensional image and the real world.

Specifically, the spatial relationship is divided into two parts (refer to Figure 3): *bbox relations* and *depth relations*. We denote *bbox relations* between two objects i and j as $\mathbf{r}^{bbox}_{i,j}$, which is a 4-dimensional vector of the relative position and size of the bounding boxes:

$$\mathbf{r}^{bbox}_{i,j} = \Big(\log(\frac{|x^c_i - x^c_j|}{w_i}), \log(\frac{|y^c_i - y^c_j|}{h_i}), \log(\frac{w_j}{w_i}), \log(\frac{h_j}{h_i})\Big), \tag{3}$$

where $(x^c_i, y^c_i), w_i, h_i$ are the center coordinate, width, and height of the i-th bounding box, respectively. This term measures coarse relative distance and position in 2-dimensional space [21], [22], [27]. As for *depth relation* between i-th and j-th object, we denote it as $\mathbf{r}^{depth}_{i,j}$ and formulate with dep_i, dep_j by:

$$\mathbf{r}^{depth}_{i,j} = \Big(\log(\frac{dep_j}{dep_i}), \log(\frac{w_j \cdot dep_j}{w_i \cdot dep_i}), \log(\frac{h_j \cdot dep_j}{h_i \cdot dep_i}), \frac{S^{inter}_{i,j}}{w_i \cdot h_i}\Big), \tag{4}$$

where $S^{inter}_{i,j}$ denotes the intersection area of bounding boxes i and j. Eq.4 aims to estimate the relative depth of object pairs (*i.e.* $\frac{dep_i}{dep_j}$), which can help the model to judge the visual distance from the image more accurately.

Finally, the 3D spatial relationship $\mathbf{r}^{spatial}_{i,j}$ is generated by simple concatenation of *bbox relation* and *depth relation*:

$$\mathbf{r}^{spatial}_{i,j} = [\mathbf{r}^{bbox}_{i,j}; \mathbf{r}^{depth}_{i,j}]. \tag{5}$$

We project $\mathbf{r}^{spatial}_{i,j}$ to a high-dimensional representation $\mathbf{R}^{spatial}_{i,j}$ with an FC layer followed by a ReLU activation:

$$\mathbf{R}^{spatial}_{i,j} = \mathrm{ReLU}(\mathrm{FC}(\mathbf{r}^{spatial}_{i,j})), \tag{6}$$

where $\mathbf{R}^{spatial} \in R^{N \times N \times d_s}$ denotes the latent pairwise spatial relationships, which can be incorporated into the attention mechanism and learned through end-to-end training for intra- and inter-modal relational alignment.

B. Question Representation

The question is embedded into a set of word embeddings $\mathbf{S} = \{\mathbf{s}_i\}^{M-1}_{i=0}$ and a global semantic vector $\mathbf{q}$ using BERT or GRU. These embeddings enable our model to attend to question-relevant objects or relations, so that question-guided visual reasoning can be performed over the object graph $\mathcal{G}$.

C. Encoding Module

On the basis of question and image representations, the encoding module aims at extracting question semantics (Question Encoder) and performing visual reasoning over the fully-connected object graph under the guidance of question semantics via message passing (Image Encoder), depicted in Figure 4.

Question Encoder. The classic Transformer [25] is exploited to encode the question semantics. Specifically, the question representations $\mathbf{S}$ are fed to a stack of Self-attention and Feed Forward layers, denoted as follows:

$$\hat{\mathbf{S}}^{l-1} = \mathrm{LN}(\mathbf{S}^{l-1} + \mathrm{SA}(\mathbf{S}^{l-1})), \tag{7}$$

$$\mathbf{S}^{l} = \mathrm{LN}(\hat{\mathbf{S}}^{l-1} + \mathrm{FFN}(\hat{\mathbf{S}}^{l-1})), \tag{8}$$

where LN, SA and FFN represent the layer normalization, self-attention and feed forward network respectively. $l = \{1, ..., L\}$ denotes the l-th Transformer layer. The output of L-th layer $\mathbf{S}^L$ is regarded as question representation, which is used to guide the image encoding and visual reasoning.

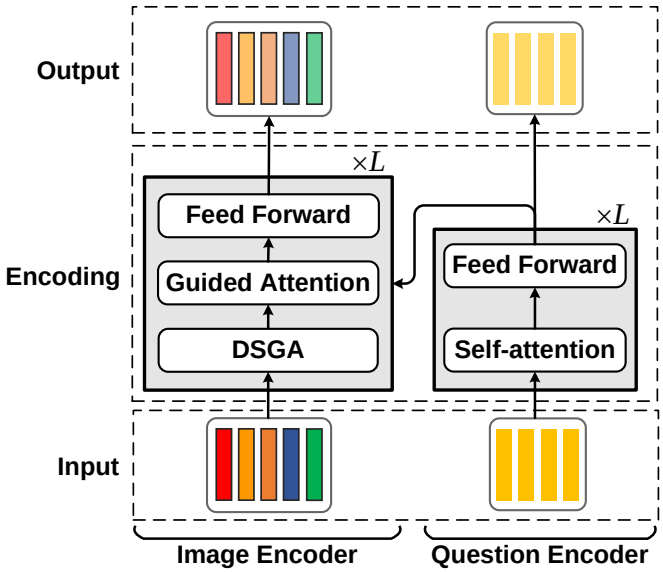

Fig. 4. Illustrations for *Encoding Module*. Transformer-based architecture is exploited to encode question semantics and object features, in which Depth-aware and Semantic Guided Relational Attention (DSGA) mechanism is incorporated into Image Encoder to facilitate the spatial visual reasoning. The residual connection and layer normalization are not displayed for simplicity.

Image Encoder. The image encoder aims to refine regional features with intra-modal and inter-modal contexts, which enables the model to perform visual reasoning among objects and attend to question-relevant regions. Specifically, the nodes features $\mathcal{V}$ are fed to a stack of DSGA, Guided Attention and Feed Forward layers, formulated as:

$$\hat{\mathbf{O}}^{l-1} = \mathrm{LN}(\mathbf{O}^{l-1} + \mathrm{DSGA}(\mathbf{O}^{l-1}, \mathbf{q})), \tag{9}$$

$$\tilde{\mathbf{O}}^{l-1} = \mathrm{LN}(\hat{\mathbf{O}}^{l-1} + \mathrm{GA}(\hat{\mathbf{O}}^{l-1}, \mathbf{S}^{L})), \tag{10}$$

$$\mathbf{O}^{l} = \mathrm{LN}(\tilde{\mathbf{O}}^{l-1} + \mathrm{FFN}(\tilde{\mathbf{O}}^{l-1})), \tag{11}$$

where $\mathbf{O} \in \mathbb{R}^{|\mathcal{V}| \times d_h}$ denotes the stacked vectors of the node features.

In Eq. (10), GA represents the guided attention (same as MCAN [9]), which refines regional features using inter-modal contexts, *i.e.,* question semantics. The difference between SA and GA is that SA derives query/key/value from regional features while GA derives key/value from question features, formulated as follows:

$$\mathrm{GA}(\mathbf{O}, \mathbf{S}) = \mathrm{softmax}(\frac{\mathbf{Q}^{O}\mathbf{K}^{S^T}}{\sqrt{d_h}})\mathbf{V}^{S}, \tag{12}$$

where $\mathbf{Q}^{O}, \mathbf{K}^{S}, \mathbf{V}^{S}$ are derived in the similar way to SA in Eq. (2), but from different inputs.

The DSGA in Eq. (10) represents our proposed Depth-aware and Semantic Guided Relational Attention mechanism, which refines regional features with intra-modal contexts, *i.e.,* visual contexts, formulated as follows:

$$\mathrm{DSGA}(\mathbf{O}, \mathbf{q}) = \mathrm{softmax}(\mathbf{A}^{DSGA})\mathbf{V}, \tag{13}$$

$$\mathbf{V} = \mathrm{FC}^{V}(\mathbf{O}), \tag{14}$$

where $\mathbf{A}^{DSGA}$ denotes the attention map, and FC is the fully-connected layer. In the vanilla Transformer, the attention map is calculated via self-attention. However, according to Eq. (2), it can be observed that the correlation scores only rely on the feature similarities, ignoring the explicit spatial structures of the image. We argue that the vanilla self-attention is insufficient for visual reasoning. Therefore, we further take the explicit spatial relations into account in DSGA mechanism. Concretely, DSGA consists of three terms: **Self-attention** (SA), **Depth-aware Attention** (DAA), and **Semantic Guided Attention** (SGA), which are combined using simple summation (depicted in Fig. (5)). Compared to SA, DAA focuses on spatially closer objects within the visual modality (intra-modal), and SGA pays more attention on question-relevant object pairs according to the semantics of language modality (inter-modal). In the following descriptions, we present the implementation of DSGA with regard to different types of image features: *objects* and *grids*.

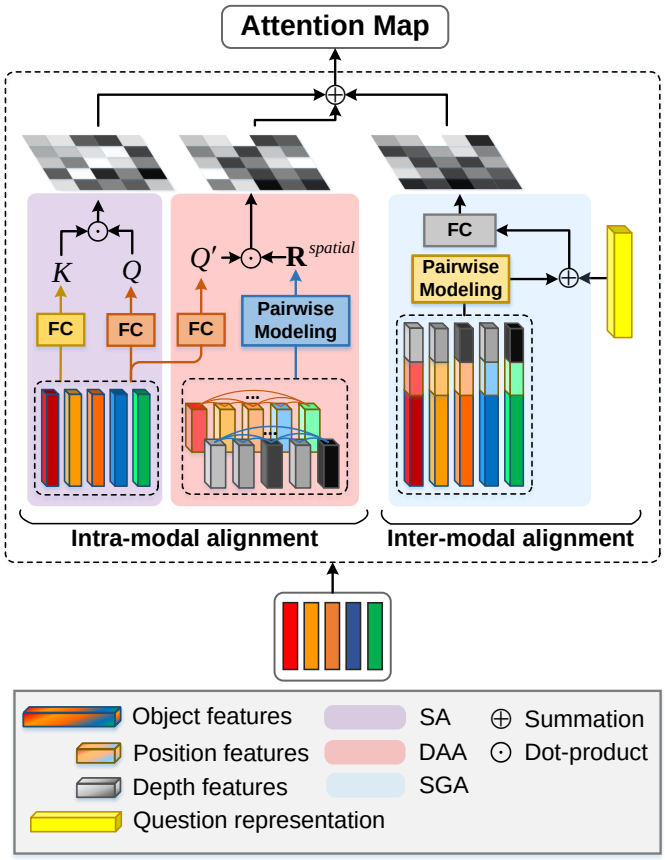

Fig. 5. Illustrations for DSGA mechanism. DSGA consists of three components: *i.e.,* self-attention (SA), semantic-guided attention (SGA), and depth-aware attention (DAA). The attention map measures correlation scores from different perspectives, *i.e.,* feature similarity, semantic relevance, and spatial relevance in the 3D space.

As for *objects*, given the input regional features $\mathbf{O} \in \mathbb{R}^{N \times d_h}$, where N denotes the number of objects, the attention map $\mathbf{A}^{DSGA}$ is calculated as follows:

$$\begin{aligned}\mathbf{A}^{DSGA} =& \mathrm{SA}(\mathbf{Q}, \mathbf{K}) + \mathrm{DAA}(\mathbf{Q}', \mathbf{K}', \mathbf{R}^{spatial}) \\ &+ \mathrm{SGA}(\mathbf{Q}'', \mathbf{K}'', \mathbf{R}^{spatial}, \mathbf{q}),\end{aligned} \tag{15}$$

where $\mathbf{Q}', \mathbf{K}' \in R^{N \times d_s}$ and $\mathbf{Q}'', \mathbf{K}'' \in R^{N \times d_s}$ are queries and keys that are calculated in the same way as $\mathbf{Q}, \mathbf{K}$ with

different projecting matrices. All the three components (SA, DAA, and SGA) output a score matrix of shape $N \times N$. For simplicity, we set the same weights for the three terms and leave models to balance the weights during training.

In Eq. 15, the first term is derived from *implicit relations*, which indicates that the correlation weights reflect the similarities of input semantics.

The second term means that the weights also rely on the spatial relations of object pairs. Following [26], the key features are omitted due to its harm on the performance, and the second term is calculated via dot-product of query and spatial relations:

$$\mathrm{DAA}(\mathbf{Q}'_i, \mathbf{K}'_j, \mathbf{R}^{spatial}_{i,j}) = \frac{\mathbf{Q}'^T_i \mathbf{R}^{spatial}_{i,j}}{\sqrt{d_h}}. \tag{16}$$

The last term in Eq. 15 indicates that the correlation weights should be adaptive to question semantics (object pairs related to question should have higher weights). Following [27], the pairwise relations are first projected to semantic space $\mathbf{R}^{sem}$, and then combined with the question semantics to calculate the attention scores, formulated as follows:

$$\mathrm{SGA}(\mathbf{R}^{sem}_{i,j}, \mathbf{q}) = \mathbf{w}^T \sigma(\mathbf{W}^a_1 \mathbf{q} + \mathbf{W}^a_2 \mathbf{R}^{sem}_{i,j}), \tag{17}$$

$$\mathbf{R}^{sem}_{i,j} = \mathbf{W}^s[\mathbf{Q}''_i; \mathbf{R}^{spatial}_{i,j}; \mathbf{K}''_j], \tag{18}$$

where $\mathbf{W}^s \in R^{d_s \times 3d_s}$ transforms the query-key pairs combined with their spatial relationships to semantic relation space. $\mathbf{W}^a_1, \mathbf{W}^a_2 \in R^{d_s \times d_s}$ are trainable weights. σ denotes the activation function, such as ReLU. $\mathbf{w}^T$ is used to obtain correlation scores. We also attempt to fuse frequency vectors into keys and values via concatenation, but get little improvement for answer prediction. For simplicity, we keep $\mathbf{R}^{sem}_{i,j}$ and remove the frequency vector term in Eq. 18.

As for *grids*, grid features retain more visual features but lead to unbearable computational complexity for pair-wise relational modeling ($\sim 600^2$). We claim that our DSGANet can generalize to grid features and obtain further performance improvement. Specifically, we generalize DSGA for grid features with DAA and omit SGA due to the computational complexity. The attention map $\mathbf{A}^{DSGA}$ is calculated by:

$$\begin{aligned}\mathbf{A}^{DSGA} =& \mathrm{SA}(\mathbf{Q}, \mathbf{K}) + \alpha \cdot [\mathrm{DAA}^{bbox}(\mathbf{Q}', \mathbf{K}') \\ &+ \mathrm{DAA}^{depth}(\mathbf{Q}', \mathbf{K}', \mathbf{dep}^{hist})],\end{aligned} \tag{19}$$

where α is a hyper-parameter to adjust the weights of two attention terms (we set $\alpha = 1/3$ in our experiments). In Eq. (19), to reduce the computational complexity, we divide DAA into two parts, *i.e.,* DAA^{bbox} and DAA^{depth}. The former part calculates the correlation scores based on 2D spatial relations, indicating that a grid might put different attention on the others with different relative coordinate distance. This term is predicted based on the query features:

$$\mathrm{DAA}^{bbox}(\mathbf{Q}', \mathbf{K}') = f^{\Delta x}(\mathbf{Q}') + f^{\Delta y}(\mathbf{Q}'), \tag{20}$$

$$f^{\Delta t}(\mathbf{Q}') = \mathbf{W}^t_2 \mathrm{ReLU}(\mathbf{W}^t_1 \mathbf{Q}'), t = \{x, y\}, \tag{21}$$

where $\Delta x, \Delta y$ denote the coordinate distance of the query-key pair in x-axis and y-axis, respectively. $\mathbf{W}^t_1 \in R^{\frac{d_h}{2} \times d_h}, \mathbf{W}^t_2 \in R^{Gs \times \frac{d_h}{2}}$ are the trainable weights. $f^{\Delta t}(\mathbf{Q}_i) \in R^{Gs}$ denotes the attention weights of i-th *grid* on the others with different relative distances, where Gs is the feature map size (H or W).

The *depth* part means that queries might put varied attention on keys with different depth. To ease the computational complexity, this term is formulated via dot-product of query-key pairs with depth features:

$$\mathrm{DAA}^{depth}(\mathbf{Q}', \mathbf{K}', \mathbf{dep}^{hist}) = \frac{\mathbf{Q}''^T \mathbf{K}''}{\sqrt{d_s}}, \tag{22}$$

$$\mathbf{Q}'' = \mathbf{W}^Q[\mathbf{Q}'; \mathbf{dep}^{hist}], \tag{23}$$

$$\mathbf{K}'' = \mathbf{W}^K[\mathbf{K}'; \mathbf{dep}^{hist}], \tag{24}$$

where $\mathbf{W}^Q, \mathbf{W}^K \in R^{d_s \times (d_k + d_h)}$ are trainable weights.

Finally, equipped with GA and our proposed DSGA mechanism in Eq. (10), the regional features are refined via intra-modal and inter-modal contexts under the guidance of question semantics. The output visual features $\mathbf{O}^L$ contains contextual information which can be used to predict the answer.

D. Prediction Module

The *Question Encoder* and *Image Encoder* separately use $L = 6$ stacked layers for image and question encoding. The outputs are further fed to an attention reduction module to obtain the attended features $\hat{\mathbf{s}}$ and $\hat{\mathbf{o}}$. Specifically, given the outputs $\mathbf{S}^L$ and $\mathbf{O}^L$ of the encoders, a self-attention mechanism is performed on $\mathbf{S}^L$ or $\mathbf{O}^L$ to obtain the aggregated representation of the whole question and image. Following [9], denoting $\mathbf{X}^L$ as an output, we adopt a two-layer MLP with ReLU activation and Dropout between them to obtain the attended feature $\hat{\mathbf{x}}$:

$$\alpha^{\mathbf{X}} = \mathrm{Softmax}(\mathbf{W}^{\mathbf{X}}_2 \mathrm{ReLU}(\mathbf{W}^{\mathbf{X}}_1 \mathbf{X}^L)), \tag{25}$$

$$\hat{\mathbf{x}} = \sum_{i=1}^{T} \alpha^{\mathbf{X}}_i \mathbf{X}^L_i, \tag{26}$$

where $\mathbf{X}$ represents $\mathbf{S}$ or $\mathbf{O}$, corresponding to the output of question and image, respectively. $\mathbf{W}^{\mathbf{X}}_2, \mathbf{W}^{\mathbf{X}}_1$ are learnable weights. T denotes the number of words or objects and $\alpha^{\mathbf{X}}$ is the normalized attention weights. $\hat{\mathbf{x}}$ denotes the final attended representation of the question or image. Then, a linear fusion function is applied to merge two representations:

$$\mathbf{h}_f = \mathrm{LayerNorm}(\mathbf{W}_o \hat{\mathbf{o}} + \mathbf{W}_s \hat{\mathbf{s}}), \tag{27}$$

$$\alpha = \mathrm{Sigmoid}(\mathbf{W}_a \mathbf{h}_f + \mathbf{b}_a), \tag{28}$$

where $\mathbf{W}_v, \mathbf{W}_q, \mathbf{W}_s, \mathbf{b}_s$ are learnable parameters, and $\alpha \in \mathbb{R}^{|\mathcal{A}|}$ denotes the probability of the $|\mathcal{A}|$ candidate answers. Following [9], the model is trained using the binary cross-entropy loss.

V. Experiments

A. Datasets

The reported results in the following sections are evaluated on the widely used VQA v2.0 [48] and GQA [49] datasets.

VQA-v2 is the most commonly used VQA benchmark dataset. It contains images from MS-COCO [50] and annotated question-answer pairs. Each image has an average of 3 questions. Each question has 10 answers annotated by different

TABLE II
STATISTICS OF SAMPLES IN VQA-V2 DATASET

Split	#Images	#Questions	#Answers
Train	82,783	443,757	4,437,570
Val	40,504	214,354	2,143,540
Test	81,434	447,793	\
All	204,721	1,105,904	\

TABLE III
STATISTICS OF BALANCE-SPLIT IN GQA DATASET

Split	#Images	#Questions	#Vocab
Train	72,140	943,000	3,097
Val	10,234	132,062	
Test-dev	398	12,578	
Test	2,987	95,336	
All	85,759	1,182,976	3,097

annotators, and the answer with most frequency is regarded as the ground-truth answer. All answers are divided into three types, *i.e., Yes/No*, *Number*, and *Other*. The dataset is split into *train*, *val* and *test* sets, and there are two test subsets to evaluate model performance online, *i.e., test-dev* and *test-std*. The statistical details are depicted in Table II. The evaluation metric (*i.e.,* accuracy) on this dataset is robust to inter-human variability, calculated by:

$$\text{Acc}(\textbf{ans}) = \min\Big\{\frac{\#\text{humans that said } \textbf{ans}}{3}, 1\Big\} \tag{29}$$

GQA is a newly proposed VQA dataset, featuring compositional questions over real-world images. It is designed to provide accurate indication of visual understanding capacity and mitigate the language priors that exist widely in previous VQA datasets. In contrast to VQA-v2 dataset, GQA is generated by leveraging Visual Genome [46] scene graph structures to create diverse reasoning questions with less language bias. Therefore, it requires more complicated reasoning skills to answer the questions. GQA consists of two splits (*i.e., balance-split* and *all-split*). The *balanced-split* consists of QA pairs with re-sampled question-answer distribution. Following the common practice, we use *balanced-split* for training and evaluation. The dataset is split into 70% train, 10% validation, 10% test and 10% challenge. The statistical details are depicted in Table III.

B. Implementation Details

The question is tokenized and encoded via BERT[1] to generate contextual word embeddings $\mathbf{Q} \in R^{M\times d_h}$ and question representation $\mathbf{q} \in R^{d_h}$. As for images, we adopt *object* features [1] extracted by pre-trained Faster R-CNN, denoted as $\mathbf{V} \in R^{100\times 2048}$, and *grid* features [2] extracted from X-152[2], denoted as $\mathbf{V} \in R^{HW\times 2048}$, where H, W denote the height and width of the feature map. The pixel-wise depth features are extracted via BTS [37] pre-trained on NYU-Depth V2 [51]. The classic Transformer-based model (*i.e.,* MCAN with layers of $L = 6$) is chosen as our base model. The hidden size d_h and spatial dim d_s are set to 768 and 160, respectively.

For model training, we use Adamax optimizer with initial learning rate of $5e^{-5}$ and warmup [52] strategy. Concretely, the learning rate starts with $5e^{-6}$ and linearly increases to $5e^{-5}$ until epoch 4. The model is trained for 13 epochs with batch-size 64 totally and the learning rate is decayed by 0.2 in epoch 11 and 13.

Baselines. For VQA v2.0 dataset, our model is compared with various state-of-the-art approaches, including MCAN [2], [9], DC-GCN [22], CMR [53], and MN-GMN [23]. A few pre-training models are also taken into account, *i.e.,* VL-BERT [10], LXMERT [11], and UNITER [12].

- **ReGAT.** The graph attention network (GAT) with relational attention mechanism. This model builds graphs of objects with three types of relations, *i.e., implicit relations*, *bbox relations* and *semantic relations*, which is a classic graph-based image encoding network.
- **DC-GCN.** The dual channel graph convolution network (GCN). This model considers the dependency relations between question words and adopts GCN to capture such dependency semantics.
- **MCAN** and **MCAN-Grid.** The Transformer-based co-attention network with different types of image features, *i.e., object* and *grid*, respectively. These two models stack Transformer layers to facilitate dense interactions between each pair of input entities from vision and language modality.
- **CMR.** The cross-modal relevance model that uses convolution layers to capture the relevance patterns of the visual and language features. The input embeddings are derived from the pre-trained LXMERT.
- **MN-GMN.** A graph-based network that incorporates both regional textual captions and visual features for image encoding. The visual/textual information are iteratively computed and updated to an external spatial memory, achieving object relational reasoning.
- **VL-BERT**, **LXMERT**, and **UNITER.** The large-scale multi-modal pre-training models based on Transformer. These models exploit large amount of unlabeled multi-modal data for pre-training, and fine-tuning in the downstream tasks to boost the performance.

For GQA, several baselines proposed by [20] and TRRNet [20] are listed for comparison. For fairness, LXMERT [11] and NSM [54] is not listed due to its usage of large-scale extra data or high-quality scene graphs.

- **LCGN.** A graph-based method that iteratively updates objects representations using GAT under the guidance of language semantics, which can be regarded as reasoning process.
- **TRRNet.** The tiered relation reasoning network that selects question-related objects and generates pairwise relations to achieve relational reasoning. This model uses reinforcement learning to determine the number of reasoning steps.

[1] *bert-base-uncased*: https://github.com/huggingface/transformers
[2] https://github.com/facebookresearch/grid-feats-vqa

TABLE IV
RESULTS ON THE VQA V2.0 DATASET.

Model	Test-dev				Test-std
	All	*Y/N*	*Num*	*Other*	*All*
Bottom-up	65.3	81.8	44.2	56.1	65.7
MFH	68.8	84.3	50.7	60.5	-
BAN	70.0	85.4	54.0	60.5	70.4
ReGAT	70.3	86.1	54.4	60.3	70.6
MCAN	70.6	86.8	53.3	60.7	70.9
CMR	72.6	-	-	-	72.6
MCAN-Grid	72.6	88.5	55.7	62.9	72.7
DC-GCN	71.2	87.3	53.8	61.5	71.5
MN-GMN	73.2	88.2	56.0	64.2	73.5
VL-BERT	70.5	-	-	-	70.8
LXMERT	72.5	-	-	-	72.5
UNITER-L	73.8	-	-	-	74.0
DSGANet-O	71.7	87.9	55.2	61.6	72.0
DSGANet-G	**74.2**	**89.9**	**58.6**	**64.9**	**74.6**

TABLE V
RESULTS ON THE GQA DATASET.

Model	Binary	Open	Accuracy
Bottom-up	66.64	34.83	49.74
MAC	71.23	38.91	54.06
BAN	76.00	40.41	57.10
GRN	74.93	41.24	57.04
LCGN	73.77	42.33	57.07
TRRNet	-	-	57.86
DSGANet-O	76.34	42.47	**58.32**

C. Experimental Results

The test results of the state-of-the-art models and our DSGANet variants on VQA v2.0 dataset are shown in Table IV. DSGANet-O and DSGANet-G denote our DSGANet model with visual features of objects and grids, respectively. Specifically, with the object features, our DSGANet-O achieves a competitive result of 71.7% overall accuracy on *test-dev*. When replacing objects with grids features, DSGANet-G significantly outperforms the current approaches on both *overall* and *per-type* accuracies. In particular, with the same visual features, our DSGANet-G outperforms MCAN-Grid by 1.6% and 1.9% on *test-dev* and *test-std*, respectively. Table V shows the comparisons on GQA dataset to evaluate our model on more complex questions. It is worth mentioning that we report *testdev* accuracy from TRRNet for fair comparison. As illustrated in Table V, our DSGANet achieves accuracy of 58.32% on *testdev* and outperforms TRRNet by 0.46%.

D. Ablation Studies

In this section, a series of ablations are conducted on VQA v2.0 and GQA to analyze the effectiveness of our proposed methods. Moreover, a subset *Rel* is split from *val* set to assess the performance on relation-related questions. Specifically, we first keep the top-100 relationships from Visual Genome dataset [46] and filter the words or phrases that are unrelated to relations, *e.g., is*, *am*, *are*, *have been*, *etc.*. Then, all the questions that contain these relationship words or phrases are split as the subset *Rel.*. The resulting subset consists of 89k question-answer pairs in total. For a fair assessment of proposed modules, we exploit MCAN as base architecture and keep the experimental settings consistent throughout the comparisons.

TABLE VI
THE EFFECTS OF DIFFERENT RELATION TYPES (*bbox* AND *depth*) ON THE VQA 2.0 VALIDATION SET.

#	DAA-Bbox	DAA-Depth	Overall	Rel.
1			68.07	65.43
2	✓		68.39	65.62
3		✓	68.32	65.55
4	✓	✓	**68.47**	**65.67**

Different relationship types. Table VI shows the ablations of two relation types (*i.e., bbox* and *depth*). Specifically, we separately exploit 4-dimensional *bbox relation* $\mathbf{r}^{bbox}$ and *depth relation* $\mathbf{r}^{depth}$ as spatial relationships for experiments. Table VI shows that either **Bbox** or **Depth** can improve the results. Specifically, compared to the base model, the incorporated *bbox* and *depth* increase the *overall* accuracy by 0.32% and 0.25%, respectively. In particular, our complete DAA improves the accuracy from 68.07% to 68.47%. We further perform additional 3 runs of our best-performing model (*i.e.,* line 4), and obtain a standard deviation of 0.04% and 0.03% on *Overall* and *Rel.*, respectively.

Effects of main components. Table VII shows the ablation studies on visual features, question encoder and attention mechanism(*i.e.,* DAA and SGA). As for object features, line 1&2 shows that Bert encoding contributes to a significant gain of +0.6% and +0.7% for *Overall* and *Rel.*, respectively. Line 2-5 illustrates the improvement via adding attention mechanisms. Specifically, the incorporated DAA and SGA increase the *overall* accuracy by 0.40% and 0.31%, respectively (line 2 *vs* 3, line 2 *vs* 4). However, little improvement is achieved between line 5 and line 3&4. The reason might be specified by the small proportion of the complex questions in VQA v2.0 dataset. Hence, we conduct additive ablation studies on GQA dataset in Table VIII. As for grid features, the consistent performance gains in Table VII prove the generalization of attention mechanism shift from *object* to *grid* input. Specifically, our full model (line 8) with grid feature achieved the highest accuracy of 70.60% and 67.95% for *overall* and *Rel.*, respectively. To measure the confidence of the results, we have performed additional 3 runs of our best-performing model over both *object* and *grid* features (*i.e.,* line 5 and 8 in Table VII), getting standard deviations of 0.08% and 0.10%, respectively.

Different datasets. To illustrate the effects of our proposed model on more complex questions, we further conduct ablations on GQA datasets. As suggested by [49], the model is trained on *train+val* and evaluated on *testdev*. Table VIII shows the testdev performance on GQA dataset. In the experiments,

TABLE VII
THE EFFECTS OF QUESTION ENCODER AND ATTENTION MODULES (DAA AND SGA) ON VQA 2.0 VALIDATION SET.

#	V-features	Bert	DAA	SGA	Overall	Rel.
1	Objects				67.43	64.68
2		✓			68.07	65.43
3		✓	✓		**68.47**	65.67
4		✓		✓	68.38	65.65
5		✓	✓	✓	**68.47**	**65.70**
6	Grids			-	69.54	67.00
7		✓		-	70.19	67.70
8		✓	✓	-	**70.60**	**67.95**

TABLE VIII
THE EFFECTS OF QUESTION ENCODER AND ATTENTION MODULES (DAA AND SGA) ON THE GQA DATASET.

#	Model	Bert	DAA	SGA	Acc.
1	Baseline				56.75
2	Our Model		✓		57.35
3			✓	✓	57.76
4		✓	✓	✓	**58.54**

we use object features by default for simplicity. As table shows, consistent improvement can be observed by adding DAA or SGA module to our base model (line 2 *vs.* line 1, and line 3 *vs.* line 1), which demonstrates the generalization of our proposed methods cross different datasets. The full model achieves the highest gain of +1.79% compared to the base model (line 1). Additionally, we perform 3 runs for the best performing model (*i.e.,* line 4 in Table VIII), and obtain a standard deviation of 1.01%.

Analysis on accuracy improvement. From Table VII and VIII, it can be observed that our proposed methods bring more significant improvement for GQA than VQA v2.0 dataset, *e.g.,* +0.41% *vs.* +0.03% by adding SGA module. This is mainly due to the different proportions of the relation-related data. According to our statistics, the proportions of such data in GQA and VQA 2.0 datasets are ∼50% and 15%∼25%, respectively. In addition, we collect relation-related words that frequently appear in VQA v2.0 dataset using POS tagging tools, and found that most of the relation words are relatively simple, *e.g.,* have, in, on, *etc.* This makes it easier for the model to solve such questions without exploiting depth information, and thus the improvements on VQA v2.0 dataset appear not obvious. Moreover, from the perspective of the image, the improvement also depends on the visual content, and most of the images on VQA v2.0 only contain one or two core objects referred by questions, which may not require relation modeling to locate the target and answer the questions.

E. Visualization

To qualitatively illustrate the effects of our proposed methods, we visualize the learned attentions from our DSGANet in Figure 7. Moreover, we conduct case studies and visualize the attention flow between objects (Figure 6 and 8), illustrating the effects of different attention modules.

In Figure 7, DSGANet-G model is exploited for visualization. Specifically, we display the attention weights of *woman* object (red solid bbox) and visualize it on the image. To obtain the weights, we split the object bounding box into grids and sum up the corresponding normalized attention values. For better visualization, the weights are smoothed with Gaussian kernel and overlaid on the image. It can be observed that in Layer-1, SA puts attention on two *women* (the middle and the right side) while DAA concentrates on the object edges. This reveals that SA gathers context mainly based on feature similarity while DAA can focus on the junction area between objects. In Layer-6, both SA and DAA are more focused and put much attention on the *clothes*. This can be explained by the fact that after multiple layers of encoding, the model has gained much contextual information and is able to locate the ground-truth area.

In Figure 6 and 8, we exploit DSGANet-O model for cases study, illustrating the effects of *bbox* and *depth* relations. Specifically, **Base**, **Bbox** and **Depth** correspond to our DSGANet variants with different relation types, *i.e., implicit relations*, *bbox relations* and *depth relations*, respectively. The three blocks display the inputs, histograms of the corresponding areas, and the visualizations of the query-key pairs from the last layer of our DSGANet. Each connection indicates a query-key pair (from *yellow* to *red* star) from the attention map, and the query is fixed to denote the object mentioned by questions (*i.e., he*, *field*). Three connections with the highest scores (summed over all heads) in each attention module are displayed in the image.

As measured by Equation 3 and 4, *bbox* relations calculate the distance via the width/height ratio of object pairs (*i.e.,* $\log(\frac{w_j}{w_i}), \log(\frac{h_j}{h_i})$). Intuitively, objects that are closer to the "camera" have larger sizes, and thus the same objects that are closer should have similar sizes. From the perspective of bboxes, the distances between objects *1-6* can be inferred by solely *bbox* relations, *i.e.,* objects *1-3* are closer to object *0* compared to objects *4-6*. Additionally, the distances can also be inferred by the depth. As shown in the second column of Figure 6, take the histogram peak value of the object *0* as the basis, objects *1-3* have closer depth values than objects *4-6*, which indicates the closer distances between objects *1-3* and *0*. In this case, benefit from the *bbox* and *depth* relations, DAA and SGA consider the spatial relations between objects, hence *looking for* spatially-relevant regions (*i.e.,* objects *1-3*) corresponding to the question.

Figure 8 shows another case, in which *bbox* relations cannot reflect the distance between objects. Specifically, from the perspective of bboxes, object *2* is closer to *0* than object *1* due to the more similar width/height ratio/size, which misleads the model focuses on the bench instead of skateboard. In this case, the main reason for this problem is the size variance between different types of objects, *e.g.,* the skateboard itself is relatively small, making it look father in distance. Therefore, *depth* information is necessary to measure more accurate relative distances between objects. As shown in the second column of Figure 8, the depth histogram is able to reflect such relative

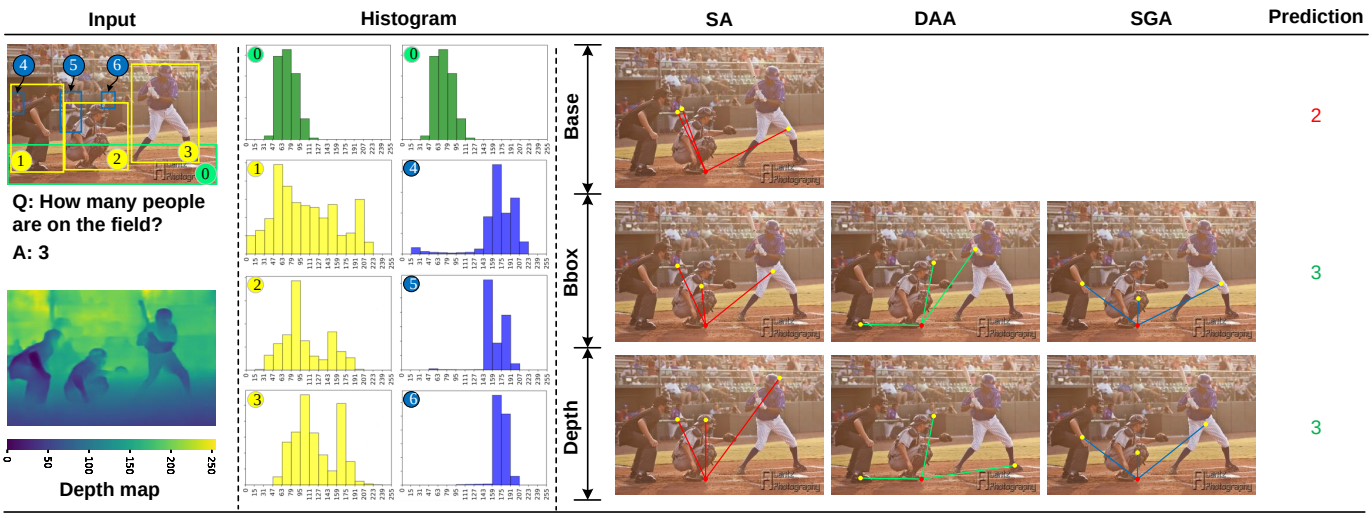

Fig. 6. Case study from our DSGANet variants and illustrations of the importance of explicit spatial relations, *i.e., bbox* and *depth* relations.

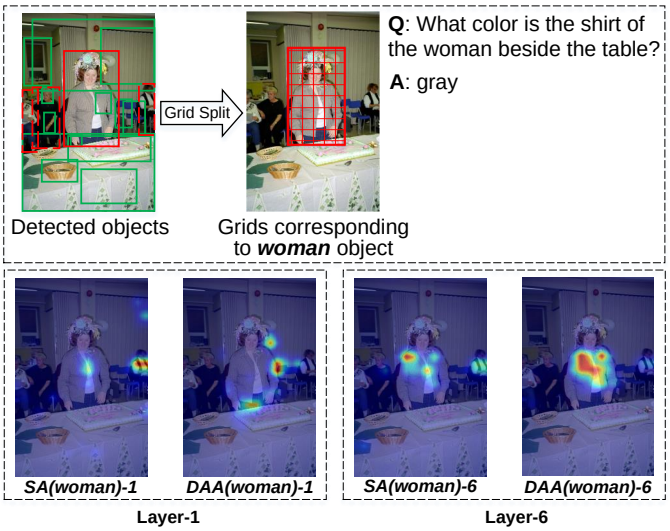

Fig. 7. Visualizations of attention weights of the grids corresponding to *woman* object (softmax($\mathbf{q}_{woman}^T \mathbf{K}/\sqrt{d}$)). SA and DAA denote Self-attention and Depth-aware Attention, respectively.

spatial relations via the margin of histogram peak values.

VI. Conclusion

We propose to construct 3D explicit spatial relations via incorporating 1D depth information into 2D bounding-box relations. On the top of the spatial relations, we develop Depth-aware and Semantic Guided Relational Attention Network (DSGANet) that refines the visual features based on both semantic and spatial relations. Besides, we generalize the attention mechanism from object to grid features and obtain consistent performance improvement. We conduct extensive experiments on VQA v2.0 and GQA datasets to evaluate the performance of our proposed DSGANet. The experimental results demonstrate that our proposed models achieve competitive performance compared to pretrained and non-pretrained models over different datasets, and the visualization of attentions provides the intuitive explanation of model behavior.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No.61602197, Grant No.L1924068, Grant No.61772076, in part by CCF-AFSG Research Fund under Grant No.RF20210005, in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018.

[2] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, "In defense of grid features for visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 267–10 276.

[3] R. Cadène, H. Ben-younes, M. Cord, and N. Thome, "Murel: Multimodal relational reasoning for visual question answering," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1989–1998, 2019.

[4] 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 *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 457–468.

[5] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," *2017 IEEE International Conference on Computer Vision*, pp. 1839–1848, 2017.

[6] H. Ben-younes, R. Cadène, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," *2017 IEEE International Conference on Computer Vision*, pp. 2631–2639, 2017.

[7] D.-K. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6087–6096, 2018.

[8] F. Liu, J. Liu, Z. Fang, R. Hong, and H. Lu, "Visual question answering with dense inter- and intra-modality interactions," *IEEE Transactions on Multimedia*, vol. 23, pp. 3518–3529, 2021.

[9] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6274–6283, 2019.

[10] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," in *International Conference on Learning Representations*, 2019.

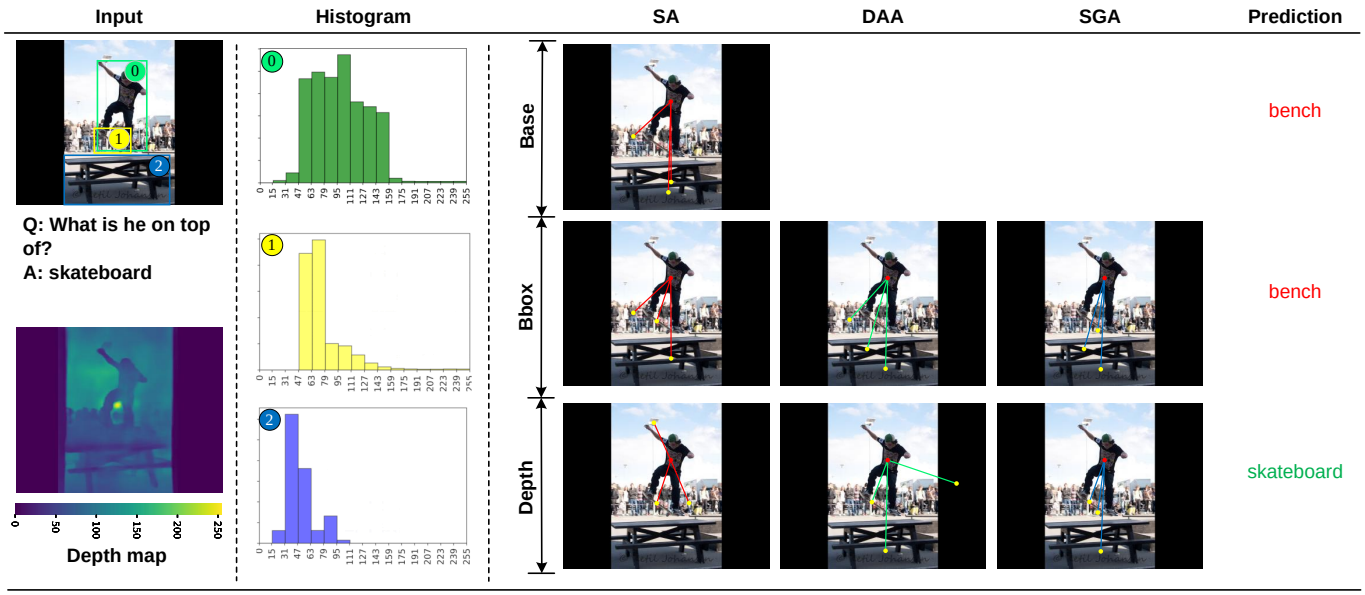

Fig. 8. Case study from our DSGANet variants and illustrations of the importance of *depth* relations when 2D bounding boxes (*i.e., bbox* relations) cannot reflect the distance in the real world.

[11] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 5100–5111.

[12] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[13] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European Conference on Computer Vision*. Springer, 2020, pp. 121–137.

[14] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5579–5588.

[15] R. Hu, J. Andreas, T. Darrell, and K. Saenko, "Explainable neural computation via stack neural module networks," in *Proceedings of the European conference on computer vision*, 2018, pp. 53–69.

[16] Z. Huasong, J. Chen, C. Shen, H. Zhang, J. Huang, and X. Hua, "Self-adaptive neural module transformer for visual question answering," *IEEE Transactions on Multimedia*, vol. 23, pp. 1264–1273, 2021.

[17] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," in *International Conference on Learning Representations*, 2018.

[18] Q. Cao, X. Liang, B. Li, and L. Lin, "Interpretable visual question answering by reasoning on dependency trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 887–901, 2021.

[19] R. Tang and C. Ma, "Interpretable neural computation for real-world compositional visual question answering," in *Chinese Conference on Pattern Recognition and Computer Vision*. Springer, 2020, pp. 89–101.

[20] X. Yang, G. Lin, F. Lv, and F. Liu, "Trrnet: Tiered relation reasoning for compositional visual question answering," in *European Conference on Computer Vision*. Springer, 2020, pp. 414–430.

[21] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," *2019 IEEE/CVF International Conference on Computer Vision*, pp. 10 312–10 321, 2019.

[22] Q. Huang, J. Wei, Y. Cai, C. Zheng, J. Chen, H.-f. Leung, and Q. Li, "Aligned dual channel graph convolutional network for visual question answering," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 7166–7176.

[23] M. Khademi, "Multimodal neural graph memory networks for visual question answering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7177–7188.

[24] R. Hu, A. Rohrbach, T. Darrell, and K. Saenko, "Language-conditioned graph networks for relational reasoning," *2019 IEEE/CVF International Conference on Computer Vision*, pp. 10 293–10 302, 2019.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] L. Guo, J. Liu, X. Zhu, P. Yao, S. chen Lu, and H. Lu, "Normalized and geometry-aware self-attention network for image captioning," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10 324–10 333, 2020.

[27] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, and A. V. D. Hengel, "Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1960–1968, 2019.

[28] H. Ben-Younes, R. Cadene, N. Thome, and M. Cord, "Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8102–8109.

[29] P. Huang, J. Huang, Y. Guo, M. Qiao, and Y. Zhu, "Multi-grained attention with object-level grounding for visual question answering," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3595–3600.

[30] Y. Liang, Y. Bai, W. Zhang, X. Qian, L. Zhu, and T. Mei, "Vrr-vg: Refocusing visually-relevant relationships," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 403–10 412.

[31] Z. Liao, Q. Huang, Y. Liang, M. Fu, Y. Cai, and Q. Li, "Scene graph with 3d information for change captioning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5074–5082.

[32] Y. Kant, D. Batra, P. Anderson, A. Schwing, D. Parikh, J. Lu, and H. Agrawal, "Spatially aware multimodal transformers for textvqa," in *European Conference on Computer Vision*. Springer, 2020, pp. 715–732.

[33] J. Yu, W. Zhang, Y. Lu, Z. Qin, Y. Hu, J. Tan, and Q. Wu, "Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 22, pp. 3196–3209, 2020.

[34] Z. Yang, Z. Qin, J. Yu, and T. Wan, "Prior visual relationship reasoning for visual question answering," *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 1411–1415, 2020.

[35] Z. Yu, Y. Cui, J. Yu, M. Wang, D. Tao, and Q. Tian, "Deep multimodal neural architecture search," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3743–3752.

[36] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3828–3838.

[37] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *arXiv preprint arXiv:1907.10326*, 2019.

[38] L. Hoyer, D. Dai, Y. Chen, A. Koring, S. Saha, and L. Van Gool, "Three ways to improve semantic segmentation with self-supervised depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 130–11 140.

[39] B. N. Patro, D. Srivastav, and V. P. Namboodiri, "Look deeper count richer: Depth based graph relation network for vqa," *Available at SSRN 4063413*.

[40] P. Banerjee, T. Gokhale, Y. Yang, and C. Baral, "Weakly supervised relative spatial reasoning for visual question answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1908–1918.

[41] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5575–5584, 2021.

[42] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A deep learning approach to visual question answering," *International Journal of Computer Vision*, vol. 125, pp. 110–135, 2017.

[43] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 30–38, 2016.

[44] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[45] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[46] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, pp. 32–73, 2016.

[47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[48] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[49] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6693–6702, 2019.

[50] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[51] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[52] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[53] C. Zheng, Q. Guo, and P. Kordjamshidi, "Cross-modality relevance for reasoning on language and vision," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7642–7651.

[54] D. A. Hudson and C. D. Manning, "Learning by abstraction: the neural state machine," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 5903–5916.

Yuhang Liu received the B.S. degree in School of Artificial Intelligence and Automation from Huazhong University of Science and Technology, Wuhan, China in 2020. He is currently working toward the B.S. degree in School of Computer Science and Technology with Huazhong University of Science and Technology, Wuhan, China. His research interests include Multimodal Representation Learning, Visual Reasoning and Visual Question Answering.

Wei Wei received the PhD degree from the Huazhong University of Science and Technology, Wuhan, China, in 2012. He is currently an Associate Professor with the Department of Computer of Science and Technology, Huazhong University of Science and Technology. He was a research fellow with Nanyang Technological University, Singapore, and Singapore Management University, Singapore. His current research interests include information retrieval, natural language processing, social computing and recommendation, cross-modal/multimodal computing, deep learning, machine learning and artificial intelligence.

Daowan Peng received his B.S. and M.S. degrees from Chongqing University of Posts and Telecommunications in China in 2017 and 2020, respectively. He is currently a Ph.D. candidate at the School of Computer Science and Technology at Huazhong University of Science and Technology. His research interests mainly include machine learning and multimodal machine learning.

Xianling Mao received the Ph.D. science degree from Peking University, China, in 2012. He is currently an Associate Professor with the Department of Computer Science and Technology, Beijing Institute of Technology, China. His major research interests include deep learning, machine learning, information retrieval, natural language processing, artificial intelligence and network data mining.

Pan Zhou is currently a full professor and PhD advisor with Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology (HUST), Wuhan, P.R. China. He received his Ph.D. in the School of Electrical and Computer Engineering at the Georgia Institute of Technology (Georgia Tech) in 2011, Atlanta, USA. He received the "Rising Star in Science and Technology of HUST" in 2017, and the "Best Scientific Paper Award" in the 25th International Conference on Pattern Recognition (ICPR 2020). He is currently an associate editor of IEEE Transactions on Network Science and Engineering. His current research interest includes: security and privacy, big data analytics, machine learning, and information networks.

APPENDIX

This appendix provides the additive illustrations about our proposed methods, *i.e.,* derivation of the *bbox* term in DAA for grid features.

A. Derivation of DAA for grid

We claim that the *bbox* term of depth-aware attention (DAA) (*i.e.,* Equation 20) for *grid* features is equivalent to that for *object* features under some assumptions. In the following description, we demonstrate how to derive Equation 20 from the perspective of bounding box relations in *object* features.

Assuming that only bounding box features are used for explicit relational modeling, the spatial relations between objects can be calculated by:

$$\mathbf{r}_{i,j}^{bbox} = \Big(\log(\frac{|x_i^c - x_j^c|}{w_i}), \log(\frac{|y_i^c - y_j^c|}{h_i}), \log(\frac{w_j}{w_i}), \log(\frac{h_j}{h_i}) \Big), \tag{30}$$

Thereafter, the spatial relation $\mathbf{R}_{i,j}^{spatial}$ is generated by an FC layer followed by ReLU activation:

$$\mathbf{R}_{i,j}^{spatial} = \text{ReLU}(\text{FC}(\mathbf{r}_{i,j}^{bbox})), \tag{31}$$

After that, we calculate the spatial attention through dot-product of the query and the spatial relation:

$$\text{DAA}^{bbox}(\mathbf{Q}'_i, \mathbf{R}_{i,j}^{spatial}) = \mathbf{Q'}_i^T \mathbf{R}_{i,j}^{spatial}. \tag{32}$$

Derivation. Suppose each grid corresponds to a square region of the original image, then the *bounding box* of grid i, j can be denoted as (x_i, y_i, x_i+1, y_i+1) and (x_j, y_j, x_j+1, y_j+1), respectively. $(x_i, y_i), (x_j, y_j)$ denotes the top-left coordinates. Thereby, Equation 30 should be rewritten as:

$$\begin{aligned} \mathbf{r}_{i,j}^{bbox} &= \Big(\log(|x_i - x_j|), \log(|y_i - y_j|), 0, 0 \Big) \\ &= \Big(\log(|\Delta x_{ij}|), \log(|\Delta y_{ij}|), 0, 0 \Big), \end{aligned} \tag{33}$$

where $\Delta x_{ij}, \Delta y_{ij}$ denotes the relative distance of grid i, j in x-axis and y-axis. Rewrite FC in Equation 31 in vector form as follows:

$$\begin{aligned} \mathbf{R}_{i,j}^{spatial} &= \text{ReLU}(\text{FC}(\mathbf{r}_{i,j}^{bbox})) \\ &= \text{ReLU}(\mathbf{W}\mathbf{r}_{i,j}^{bbox}) \\ &= \text{ReLU}([\mathbf{w_0}\ \mathbf{w_1}\ \mathbf{w_2}\ \mathbf{w_3}][\log(|\Delta x_{ij}|)\ \ \log(|\Delta y_{ij}|)\ 0\ 0]^T) \\ &= \text{ReLU}(\mathbf{w_0} \cdot \log(|\Delta x_{ij}|) + \mathbf{w_1} \cdot \log(|\Delta y_{ij}|)), \end{aligned} \tag{34}$$

Since $\Delta x, \Delta y$ is discrete for grid features, we initialize a trainable vector $\mathbf{w}_{\Delta x}^X, \mathbf{w}_{\Delta y}^Y$ randomly for each value of Δx or Δy, which indicates the relative distance embedding, hence $\mathbf{R}_{i,j}^{spatial}$ is rewritten as follows:

$$\mathbf{R}_{i,j}^{spatial} = \text{ReLU}(\mathbf{w}_{\Delta x_{ij}}^X + \mathbf{w}_{\Delta y_{ij}}^Y), \tag{35}$$

Assume that the effects of the relative distance on x-axis and y-axis are separate:

$$\begin{aligned} \mathbf{R}_{i,j}^{spatial} &= \text{ReLU}(\mathbf{w}_{\Delta x}^X) + \text{ReLU}(\mathbf{w}_{\Delta y}^Y) \\ &= \mathbf{w'}_{\Delta x_{ij}}^X + \mathbf{w'}_{\Delta y_{ij}}^Y, \end{aligned} \tag{36}$$

where $\mathbf{w'}_*^*$ denotes another initialized non-negative vector that satisfy $\mathbf{w'}_*^* = \text{ReLU}(\mathbf{w}_*^*)$. Take this into account, Equation 32 is rewritten as follows:

$$\begin{aligned} \text{DAA}^{bbox}(\mathbf{Q}'_i, \mathbf{R}_{i,j}^{spatial}) &= \mathbf{Q'}_i^T \mathbf{R}_{i,j}^{spatial} \\ &= \mathbf{Q'}_i^T (\mathbf{w'}_{\Delta x_{ij}}^X + \mathbf{w'}_{\Delta y_{ij}}^Y) \\ &= \mathbf{Q'}_i^T \mathbf{w'}_{\Delta x_{ij}}^X + \mathbf{Q'}_i^T \mathbf{w'}_{\Delta y_{ij}}^Y, \end{aligned} \tag{37}$$

We pack the weights $\mathbf{w'}_{\Delta x}^X, \mathbf{w'}_{\Delta y}^Y$ corresponding to all Δx and Δy values into matrices $\mathbf{W'_X}, \mathbf{W'_Y}$, and get the equation as follows:

$$\begin{aligned} \text{DAA}_{i,j}^{bbox}(\mathbf{Q}'_i, \mathbf{R}_{i,j}^{spatial}) &= (\mathbf{W'_X}^T \mathbf{Q'}_i)[\Delta x_{ij}] + (\mathbf{W'_Y}^T \mathbf{Q'}_i)[\Delta y_{ij}] \\ &= f^{\Delta x}(\mathbf{Q})[\Delta x_{ij}] + f^{\Delta y}(\mathbf{Q})[\Delta y_{ij}], \end{aligned} \tag{38}$$

where $f^{\Delta x}(\cdot), f^{\Delta y}(\cdot)$ denotes FC layers with different weights. In our implementation, we adopt multi-layer perceptron (MLP) for expressive ability.