Weakly-Supervised 3D Spatial Reasoning for Text-Based Visual Question Answering

Hao Li, Jinfua Huang, Peng Jin, Graduate Student Member, IEEE, Guoli Song, Qi Wu, Member, IEEE, and Jie Chen, Member, IEEE

Abstract—Text-based Visual Question Answering (TextVQA) aims to produce correct answers for given questions about the images with multiple scene texts. In most cases, the texts naturally attach to the surface of the objects. Therefore, spatial reasoning between texts and objects is crucial in TextVQA. However, existing approaches are constrained within 2D spatial information learned from the input images and rely on transformer-based architectures to reason implicitly during the fusion process. Under this setting, these 2D spatial reasoning approaches cannot distinguish the fine-grained spatial relations between visual objects and scene texts on the same image plane, thereby impairing the interpretability and performance of TextVQA models. In this paper, we introduce 3D geometric information into the spatial reasoning process to capture the contextual knowledge of key objects step-by-step. Specifically, (i) we propose a relation prediction module for accurately locating the region of interest of critical objects; (ii) we design a depth-aware attention calibration module for calibrating the OCR tokens’ attention according to critical objects. Extensive experiments show that our method achieves state-of-the-art performance on TextVQA and ST-VQA valid split. Besides, we also verify the generalizability of our model on the text-based image captioning task.

Index Terms—Text-based visual question answering, spatial reasoning, 3D geometric information, transformer.

I. INTRODUCTION

TASKSs around scene-text have broad prospects in various areas including automatic driving and online shopping. As a representative task focusing on scene-text in images, text-based visual question answering (TextVQA) [1], [2], [3] needs to answer the given question about the image with multiple objects and scene texts. Thus, in addition to the simple object-object spatial reasoning in the VQA task [4], TextVQA models also require the ability to read textual information and reason the spatial relationship between different visual entities and texts in the image.

Existing Transformer-based TextVQA models generally adopt the 2D spatial information learned from images for spatial reasoning. Most models [5], [6] use the absolute spatial information of objects and optical character recognition (OCR) tokens to supplement visual features. They straightly add the spatial features into the visual feature as the fusion of spatial and contextual information. Some inspiring works [7], [8] also explore rule-based spatial correlations from objects’ and OCRs’ absolute coordinates. Generally, they use the vanilla transformer structure to implicitly reason for the answer during the multimodal fusion process. However, without structured model input, the transformers only show shallow reasoning ability on the reasoning tasks [9].

Therefore, existing TextVQA models cannot distinguish fine-grained spatial relations between multiple objects and scene texts in the same image plane. As shown in Fig. 1, the 2D spatial reasoning model receives the bounding box information, which only contains 2D spatial information. It leads to spuriousness when multiple object bounding boxes and OCR bounding boxes are close to each other. To this end, the model cannot distinguish the spatial relations between “Budget”, “GMC”, and “carriage”. Due to the wrong
prediction that “Budget” and “GMC” are on the same object “carriage”, “GMC” becomes the distraction leading the model to predict the wrong answer “GMC”. We find that leading 3D information into the reasoning process can overcome these issues. Fig. 2 shows that leading 3D information can boost the performance of TextVQA models.

In Fig. 3 we analyze the spatial-related cases in one VQA dataset: Visual Genome, and two Scene-Text VQA datasets: TextVQA and STVQA. Fig. 3(a) shows that the spatial-related questions accounted for more than 40% of the total questions in TextVQA, much higher than the rate in the traditional VQA task. We further make a human evaluation for the questions in TextVQA to find out the rate of the questions requiring depth information to answer. Fig. 3(b) shows that in the spatial-related question sets (“on”, “under”, “top”), more than 81% of the question require depth information to answer. As for the other question, there still have 32% of the question require depth information to answer. Hence, to achieve complex spatial reasoning on multiple objects and OCR tokens, we introduce 3D geometric information into the TextVQA reasoning model. The adoption of 3D geometric information is motivated by an intuitive observation: in most cases, the texts are attached to the surface of objects. Moreover, with the depth information, the model can better understand the real-world 3D scene inside the images [10]. Therefore, the 3D geometric information might assist the TextVQA model to understand the spatial structure of visual concepts in an image, which is extremely important for the reasoning process of transformers [11], [12]. As shown in Fig. 1, with the guidance of 3D information, our model alleviates the spuriousness problem on the same image plane and accurately predicts the fine-grained spatial relations such as “Budget-on-carriage” and “GMC-on-car”. Benefiting from correct spatial relation prediction, our model answers the question correctly.

However, it is challenging for TextVQA models to take advantage of 3D geometric information. Previous models apply the vanilla transformer structure as the fusion module, but this structure lacks spatial inductive biases and hardly preserves the 3D spatial structure of the input image [13], [14], [15], [16]. Since existing models are delicately designed for 2D spatial reasoning, they cannot comprehend 3D geometric information by directly replacing 2D spatial information.

In this paper, to facilitate the 3D understanding ability of the model, we propose the Depth-Aware TextVQA Network (DA-Net) with two novel modules: the depth-aware attention calibration module and the relation prediction module. Our depth-aware attention calibration module follows a step-by-step knowledge-transferring procedure: it first generates attention for object tokens and then transfers the object attention score for calibrating the OCR tokens’ attention under the guidance of 3D geometric information. Such a step-by-step process is similar to human reasoning [11]: humans first locate the critical objects in the image based on the given question; then they identify the OCR tokens on the detected objects and get the correct answer. To model the above gradual working patterns with 3D geometric information, our DA-Net introduces a relation prediction module to strengthen the model’s understanding of 3D spatial relationships, which is beneficial for locating critical objects.

To demonstrate the effectiveness of our DA-Net, we conduct the experiments on two widely-used datasets, the TextVQA dataset [17] and the ST-VQA dataset [2]. Extensive experiments demonstrate that our method outperforms the state-of-the-art TextVQA methods. We also split the spatial-related questions in both datasets and generate their 3D subsets. Fig. 2 shows that our model achieves state-of-the-art performance on the TextVQA task and has a remarkable improvement compared with the baseline model: SSbase (e.g., 5.7% on the TextVQA 3D subset and 12.1% on the ST-VQA 3D subset). The visualization for spatial maps and attention scores further prove our DA-Net can understand and leverage 3D spatial information. What’s more, we achieve promising performance on the text-based image captioning task, which verifies the generalization of our model.

Overall, the main contributions of our work are as follows:

- As far as we know, we are the first to introduce 3D geometric information into the TextVQA task, in order to effectively handle the spuriousness of existing 2D models' spatial reasoning process.
- With 3D spatial information, we propose a relation prediction task to strengthen the model’s understanding of 3D spatial relationships for better locating the region of interest of key objects.
- To identify the OCR tokens on the corresponding object, we design a depth-aware attention calibration module to calibrate the OCR token attention based on the key object.
- Extensive experiments show that our model achieves state-of-the-art performance on two TextVQA datasets. Results on their 3D subsets further demonstrate the spatial understanding and reasoning ability of our model.

II. RELATED WORK

A. Text-Based Visual Question Answering

As a further task of VQA [18] and VideoQA [19], TextVQA has seen rapid development in recent years. The TextVQA task aims to understand and reason scene texts in images. A model first detects OCR text and visual objects in the images and then answers related questions. LoRRA [3] extends Pythia [20] with an OCR detection module and a cross-modality attention module to reason over a combined list of answers from a static vocabulary and detected OCR tokens. M4C [21] replaces the copy mechanism in LoRRA with a dynamic pointer network and utilizes a transformer to combine multimodal information into a joint embedding space. SSbase [6] claims a simple attention mechanism can obtain comparable performance with previous sophisticated multi-modality frameworks. Recent works [22], [23], [24], [25] apply large-scale pretraining datasets on TextVQA. However, these existing models straightforwardly add a location embedding of the absolute location to...
the object and OCR features, lacking explicit spatial reasoning between different visual contents.

B. Leveraging Spatial Reasoning for TextVQA

Spatial reasoning ability is significant in behavioristics and artificial intelligence [26], [27]. In the visual reasoning domain, including VQA [28] and VideoQA [29] tasks, the region-level spatial relationship has been proved significantly helpful [30], [31], [32], [33], [34]. For TextVQA, MM-GNN [35] represents an image as a graph consisting of three sub-graphs depicting visual, semantic, and numeric modalities. Then MM-GNN uses three aggregators to guide the reasoning process. SMA [7] uses a structural graph representation to encode the object-object, object-text, and text-text relationships appearing in the image and then designs a multimodal graph attention network to reason over it. SA-M4C [8] proposes a novel spatially aware self-attention layer such that each visual entity only looks at neighboring entities defined by a spatial graph. However, these models extract the spatial relationship between different visual content from 2D bounding boxes. When extracting the crucial spatial relations for the TextVQA, such as the object-text spatial relationship, lacking 3D geometric information reduces the relation accuracy. Comparing with the existing models, our DA-Net model is a weakly-supervised spatial reasoning model using depth information to explicitly build spatial relationships.

C. Leveraging Depth Estimation in Multimedia

Monocular (single-image) depth estimation remains a challenging problem, with learning-based methods pushing the envelope [36], [37], [38]. AdaBins [39] uses a transformer-based architecture that adaptively divides depth ranges into variable-sized bins and estimates depth as a linear combination of these depth bins. It is the state-of-the-art monocular depth estimation model for outdoor and indoor scenes. We use it as the depth extractor to guide our models for better spatial reasoning. In multiple areas, only using 2D images cannot adequately guide the model with the 3D geometry [36]. Thus, the depth estimation method is leveraged across multimedia fields. For generation tasks, DaGAN [40] uses the learned depth maps to estimate sparse facial key points and generate highly realistic faces. DepthGAN [41] uses depth maps as 3D prior to assisting in synthesizing indoor scenes. For classification tasks such as object detection and visual question answering, pseudo-LiDAR-based methods [42], [43], [44], [45] lift images to 3D coordinate via monocular depth estimation. In [10], extra depth information is used as weak supervision to enhance the spatial estimation between objects. However, no existing models consider the depth correlation between objects and OCRs in TextVQA.

D. 2D Spatial Reasoning VS 3D Spatial Reasoning

2D and 3D spatial reasoning occur in many computer vision and multimedia tasks including person re-identification [46], [47], visual grounding [48], VQA [49] and VideoQA [29]. In many scenarios, 2D visual information has intrinsic limitations such as illumination, pose, expression and disguise [50]. Containing richer geometric structures, 3D visual information can provide more discriminative spatial features for the spatial reasoning process and overcome the spuriousness of visual changes. Although 3D (RGB-D) contains more abundant information than 2D (RGB), there are still challenges to applying 3D information in the spatial reasoning process. Firstly, reasoning models need to design delicately to take advantage of 3D geometric information. Traditional 2D reasoning models hardly learn 3D geometric information by directly replacing the 2D spatial information. Secondly, the acquisition of 3D information cannot be accomplished by crawling the Web like how 2D images are collected. For the TextVQA task, we propose to apply 3D spatial reasoning and also need to address the above two challenges. Compared with VQA which only needs to reason between different objects, TextVQA requires spatial reasoning on multiple objects and OCR tokens. Our DA-Net spatial reasoning focuses on building the subordination between objects and OCRs. Intuitively, we tend to build the “on” relationship between objects and OCRs. To explore the geometric structure, we design a depth-aware attention calibration module and the relation prediction auxiliary task. To obtain 3D annotations cheaply, we predict the 3D information using a well-pretrained depth estimation model instead of sampling by the depth camera.

III. DEPTH-AWARE TEXTVQA NETWORK

We propose the Depth-Aware TextVQA Network (DA-Net) for the TextVQA task, applying the 3D geometric information in the spatial reasoning procedure. Fig. 4 shows the pipeline of our DA-Net. Compared with the vanilla reasoning pipeline of TextVQA, we introduce three additional new modules. (a) Depth Estimation module extracts the relative depth map from the raw image and estimates the depth information.
Fig. 4. An illustration of the proposed Depth-Aware TextVQA Network (DA-Net). It contains three modules to exploit 3D information. (a) The depth estimation module extracts the pixel-level depth map and uses mean and centroid estimations to calculate the objects’ and OCRs’ depth. (b) The depth-aware attention calibration module (DAC) enhances the feature attention between depth-relevant OCR and key object tokens and calibrates the irrelevant OCR attention score. In addition, we combine the $f_{obj}$, $f_{OCR}$, and $Q$ as the DAC module input and perform the DAC module once to calibrate the OCR attention score. (c) The relation prediction module predicts the image’s 3D geometric structure using objects’ features. In this part, the binary cross-entropy loss between the predicted relation and GT relation helps our model understand 3D information.

for each OCR and object token. (b) Depth-aware Attention Calibration module simulates the step-by-step human-like reasoning process. It also enhances the feature attention between depth-relevant OCR and object tokens. (c) Relation Prediction Head further helps our model understand the 3D depth information and learn the implicit spatial structure from the image. Other components are introduced in Section III-A.

A. Preliminaries

General TextVQA models such as [21] and [51] mainly consist of four components: multi-modality inputs, OCR noise reduction, multi-modality transformer, and answer prediction decoder. Our DA-Net adopts the same structure as the general models for these components.

1) Multi-Modality Inputs: Given a text-related question and an image, models extract three modalities and prepare their corresponding features. Concretely, for question tokens, models use pre-trained BERT [52] to generate the question word embedding $Q = \{q_i\}_{i=1}^L$, which is 768 $\times$ $L$ dimension and $L$ is the length of the question. For OCR features, models use off-the-shelf OCR detection models [53], [54] to locate and extract OCR $= \{N_{OCR}\}_{i=1}^{50}$. For each image, we pad the OCR numbers into 50. For object visual features, models use pre-trained object detectors [55] to locate and extract Object $= \{N_{obj}\}_{i=1}^{80}$. For each image, we pad the object numbers into 80.

2) OCR Noise Reduction: The OCR tokens extracted by OCR detectors contain distractions including repeating OCR tokens and OCR subsequence tokens. Following SA-M4C [8] and LOGOs [56], we use denoising strategies to remove the distractions in OCR tokens. Specifically, we first calculate the IoU between each OCR token pair and remove the item with a high IoU score, which might be repeating OCR tokens. Then, we compare each OCR token pair and remove short subsequences tokens.

3) Multi-Modality Transformer: Most TextVQA models use the transformer as the fusion encoder of three modality inputs. For a fair comparison, we use the generative transformer [52] as our cross-modality fusion module to fuse 768 dimension vectors from OCR, Object, and question tokens.

4) Answer Prediction Decoder: M4C [21] proposes a powerful answer decoder module that iteratively decodes answers using a dynamic pointer network. Following other recent works [5], [6], [7], our model applies the same answer prediction decoder as M4C for a fair comparison.

B. Depth Estimation

1) Pixel-Level Depth Computation: To extract pixel-level depth information from images in TextVQA datasets [2], [17], we utilize an open-source monocular depth computation method AdaBins [39], which is the state-of-the-art method. AdaBins divides the depth range into bins whose center value is estimated adaptively per image. The final depth values are estimated as linear combinations of the bin centers. We obtain depth-value $d(i,j)$ for each pixel $(i,j), i \in \{1, \ldots, H\}, j \in \{1, \ldots, W\}$ in the image, where $H$ is the height of the image and $W$ is the width of the image.

2) Extracting OCR Depth: TextVQA models often use the bounding box for each OCR token in the image as
of the i key tokens, models [5], [6] use attention blocks to summarize into a joint embedding space. However, it consumes much
fuse homogeneous entities from OCRs, objects, and questions
C. Depth-Aware Attention Calibration Module

The depth of the object token is calculated as the mean depth of all points in the bounding box:

\[
d_{\text{obj}} = \frac{1}{n} \sum_{i,j} d(i, j), \quad i \in [x_e - x_c, x_c + x_e], \quad j \in [y_e - y_c, y_c + y_e].
\]

Thus our model uses 3D bounding box \([B_{\text{obj}}^i, d_{\text{obj}}^i]\) as each object token’s spatial coordinates. \(B_{\text{obj}}^i\) is the bounding box of the \(i^{th}\) object \(\text{obj}_i\).

C. Depth-Aware Attention Calibration Module

The multi-modality fusion part of TextVQA models needs to fuse homogeneous entities from OCRs, objects, and questions into a joint embedding space. However, it consumes much computation and cannot efficiently extract key token features from all entities. To filter out irrelevant tokens and highlight key tokens, models [5], [6] use attention blocks to summarize token features from objects and OCRs. However, the previous summary module pays little attention to the 3D relation between OCRs and objects.

In this paper, we extend the feature summary module to Depth-aware Attention Calibration (DAC) module. As shown in Fig. 5, it uses depth-aware weight transfer over the input OCR and object tokens. The details are as follows.

1) Question Feature: We start with the OCR tokens. To efficiently calculate the attention scores between OCR and the question sequence, we first summarize the question sequence into an individual entity. Specifically, given a question \(Q = \{q_i\}_{i=1}^L\), we use two convolution layers and one ReLU activation function between them to summarize the question sequence into an individual entity \(Q_{\text{OCR}}\):

\[
q_i^{\text{OCR}} = \text{Conv}(\text{ReLU}[\text{Conv}(q_i)]), \quad i \in [1, \ldots, L];
\]

\[
Q_{\text{OCR}} = \sum_{i=1}^L q_i \cdot \text{Softmax}(q_i^{\text{OCR}}).
\]

The \(Q_{\text{obj}}\) for object features can be calculated in the same way.

\[
q_i^{\text{obj}} = \text{Conv}(\text{ReLU}[\text{Conv}(q_i)]), \quad i \in [1, \ldots, L];
\]

\[
Q_{\text{obj}} = \sum_{i=1}^L q_i \cdot \text{Softmax}(q_i^{\text{obj}}).
\]

2) Attention Scores: We define object features \(f_{\text{obj}} = (x_1, \ldots, x_N) \in \mathbb{R}^{D \times N}\) and OCR features \(f_{\text{OCR}} = (y_1, \ldots, y_N) \in \mathbb{R}^{D \times M}\). We use \(S_{\text{obj}} = [a_1, \ldots, a_N]\) to represent the object attention score and \(S_{\text{OCR}} = [\beta_1, \ldots, \beta_M]\) to represent the OCR attention score, where:

\[
\alpha_i = \text{Softmax}\left(\frac{Q_{\text{obj}}(x_i)^\top}{\sqrt{d}}\right), \quad \forall i \in [1, \ldots, N],
\]

\[
\beta_j = \text{Softmax}\left(\frac{Q_{\text{OCR}}(y_j)^\top}{\sqrt{d}}\right), \quad \forall j \in [1, \ldots, M].
\]

3) Depth-Aware Weight Transfer: To highlight the spatial relationship between OCRs and objects, we transfer object weights to OCR weights under the supervision of 3D spatial information. Firstly, we define \(\phi_{cr}\) as the cover rate between the OCR bounding box \(B_{\text{OCR}}\) and object bounding box \(B_{\text{obj}}\).

\[
\phi_{cr}(B_{\text{obj}}^i, B_{\text{OCR}}^j) = \frac{\text{Area}(B_{\text{obj}}^i \cap B_{\text{OCR}}^j)}{\text{Area}(B_{\text{OCR}}^j)}.
\]

Secondly, we generate the \(\Delta ij\), representing the weight transfer rate from \(\text{obj}_i\) to \(\text{OCR}_j\). \(\Delta ij\) positively correlates with the cover rate between the OCR and object bounding boxes.

\[
\Delta ij = \text{Softmax}(\phi_{cr}(B_{\text{obj}}^i, B_{\text{OCR}}^j) \cdot [1 - d_{\text{obj}}^i + d_{\text{OCR}}^j]).
\]
Thirdly, the attention score of OCR \(j \) in \( S_{OCR} \) is added by the attention score of \( obj_i \) in \( S_{obj} \) with the weight transfer rate \( \Delta_{ij} \). We compute the weighted sum of \( f_{obj} \) and \( f_{OCR} \) using the updated \( S_{obj} \) and \( S_{OCR} \). \( f_{obj} \) and \( f_{OCR} \) represent the summary feature for objects and OCRs.

\[
F_{\text{obj}} = \sum S_{\text{obj}} \cdot f_{\text{obj}}, \tag{12}
\]

\[
F_{\text{OCR}} = \sum (S_{\text{OCR}} + \Delta \cdot S_{\text{obj}}) \cdot f_{\text{OCR}}. \tag{13}
\]

We combine the \( f_{\text{obj}} \), \( f_{\text{OCR}} \), and \( Q \) as the DAC module input and perform the DAC module once to calibrate the OCR attention score. After the DAC module, we put the object visual features, OCR visual and semantic features, and question features into the multi-modality transformer for cross-modality feature fusion.

### D. Relation Prediction Task

General models [5], [6] use linear layers to fuse the bounding box spatial information \( B \) into visual features \( f_{\text{obj}} \).

\[
f_{\text{obj}} = LN(W_e f_{\text{obj}}) + LN(W_b x B), \tag{14}
\]

where \( W_e \) and \( W_b \) are trainable parameters. \( LN \) represents the layer normalization. However, the spatial fusion method lacks supervision and penalty. Thus, the model tends to ignore the tokens’ spatial information.

We design a relation prediction auxiliary task to add spatial supervision to object-object and object-OCR relationships to address these limitations. For object-object pairs in the object set: \( Obj \), a relation prediction head generates an interrelation map \( R_{N \times N} \) using bounding boxes and depths in \( Obj \). As shown in Fig. 6, for \( obj_i, obj_j \in Obj \), a relation prediction head uses a two-layer feed-forward network to generate their interrelation \( r_{ij} \in Rel \).

\[
f_{\text{obj}} = LN(W_e f_{\text{obj}}) + LN(W_b x [B_{ij}, d_{ij}]), \tag{15}
\]

\[
r_{ij} = W \ast (W_{obj} f_{obj} - W_{obj} f_{obj}^i) + b, \tag{16}
\]

where \( W_{obj} \) and \( W \) are trainable parameters.

As for the object-OCR pair, take \( obj_i, OCR \) as an example, we calculate their interrelation \( r_{ij} \) in a rather similar way.

\[
f_{\text{OCR}} = LN(W_e f_{\text{OCR}}) + LN(W_b x [B_{OCR}, d_{OCR}]), \tag{17}
\]

\[
r_{ij} = W \ast (W_{obj} f_{OCR} - W_{obj} f_{obj}^i) + b, \tag{18}
\]

1) Training Loss: To evaluate the accuracy of the inter-relation map \( R \) between object-object pairs and object-OCR pairs, we generate the interrelation map \( Z_{obj} \), \( Z_{\text{OCR}} \in \mathbb{R}^{N \times N} \) as the pseudo labels of interrelationship. Specifically, \( z^{obj}_{ij} \in Z_{obj} \) is \( obj_{ij} \) and \( obj \) approximate interrelation. Similar to \( \Delta_{ij} \), \( z_{ij} \) is positively correlated with the cover rate between two object bounding boxes. \( z_{ij} \) is negatively correlated with the depth interpolation between objects. We calculate the \( z^{\text{OCR}}_{ij} \in Z_{\text{OCR}} \) for object \( i \) and OCR \( k \) in the same way.

\[
z^{obj}_{ij} = \phi_{cr}(B_{obj}, B_{obj}^i) \ast (d_{obj} - d_{obj}^i), \tag{19}
\]

\[
z^{\text{OCR}}_{ik} = \phi_{cr}(B_{\text{OCR}}, B_{OCR}^i) \ast (d_{\text{OCR}} - d_{\text{OCR}}^i). \tag{20}
\]

We use mean multi-label Binary Cross-Entropy (BCE) loss to evaluate the similarity between the pseudo-label: relation matrix \( Z^{\text{OCR}} \), \( Z^{obj} \) and the predicted relation matrix \( rel \):

\[
L_{\text{rel}} = \phi_{\text{BCE}}(Z^{\text{OCR}}, rel) + \phi_{\text{BCE}}(Z^{obj}, rel) \tag{21}
\]

**Algorithm 1 DA-Net**

```plaintext
Input: Question embedding \( Q_{OCR} \) and \( Q_{obj} \), Object and OCR features \( f_{obj} \), \( f_{OCR} \), Object and OCR bounding box \( B_{obj} \), \( B_{OCR} \), Object and OCR depth information \( d_{obj} \), \( d_{OCR} \), The multimodality transformer \( MMT \), Parameters \( \Theta(i) \) of DA-Net. The number of iterations \( i = 0 \), Learning rate \( \lambda(i) \)

1 while Iter=i to total iteration do
  1) Weight Calibration:
  2) \( S_{OCR} \leftarrow \sigma(Q_{OCR} \cdot f_{OCR}) \) by Eq. (8);
  3) \( S_{obj} \leftarrow \sigma(Q_{obj} \cdot f_{obj}) \) by Eq. (9);
  4) \( \Delta \leftarrow \sigma(CR(B_{obj}, B_{OCR}) \ast (1 - (d_{obj} - d_{OCR}))) \) by Eq. (11);
  5) \( f_{obj} \leftarrow \sum S_{obj} \cdot f_{obj} \) by Eq. (12);
  6) \( f_{OCR} \leftarrow \sum (S_{OCR} + \Delta \cdot S_{obj}) \cdot f_{OCR} \) by Eq. (13);
  7) Fusion and Prediction:
  8) \( y_{pred} = MMT(Q_{obj}, Q_{OCR}, f_{obj}, f_{OCR}) \);
  9) Update Parameters:
     Compute \( \lambda_{\text{semantic}} \), \( \lambda_{\text{spatial}} \), and gradients \( x_i \) by
     Eq. (21) and Eq. (22);
     Update parameters \( \Theta(i+1) \) by
     \( \Theta(i+1) = \Theta(i) - \lambda(i) \frac{\partial L_{\text{rel}}}{\partial x_i} \)
13 end

Output: parameters of DA-Net \( \Theta(i+1) \)
```

The ANLS metric can capture OCR accuracy and evaluate reasoning ability. The ANLS gives an intermediate score between 0.5 and 1 that will softly penalize the OCR mistakes in TextVQA [2]. Formally, the ANLS details are as follows:

\[
ANLS = \frac{1}{N} \sum_{i=0}^{N} \left( \max_j s(a_{ij}, o_{q_i}) \right)
\]

\[
s(a_{ij}, o_{q_i}) = \begin{cases} 1 - NL(a_{ij}, o_{q_i}), & \text{if } NL(a_{ij}, o_{q_i}) < \tau \\ 0, & \text{if } NL(a_{ij}, o_{q_i}) \geq \tau \end{cases}
\]

where \( N \) is the total number of questions in the dataset, \( M \) is the total number of ground-truth answers per question, \( a_{ij} \) are the ground-truth answers where \( i = \{0, \ldots, N\} \), and \( j = \{0, \ldots, M\} \), and \( o_{q_i} \) is the network’s prediction answer for the \( i \)th question \( q_i \). \( NL(a_{ij}, o_{q_i}) \) represents the normalized Levenshtein distance between \( a_{ij} \) and \( o_{q_i} \). The NL ranges from 0 to 1. \( \tau \) is a threshold that penalizes metrics larger than this value, thus the final score will be 0 if the NL is larger than \( \tau \). Following the setting of STVQA [2], we set \( \tau = 0.5 \).

To evaluate the accuracy of the answer, we follow the loss set from [6]. The semantic loss comes from two parts: the BCE loss and the ANLS loss:

\[
L_{\text{semantic}} = \phi_{\text{BCE}}(y_{pred}, y_{gt}) + \phi_{\text{ANLS}}(y_{pred}, y_{gt}),
\]

where \( y_{pred} \) is the predicted answer and \( y_{gt} \) is the ground truth answer. To this end, the total training loss is the weighted sum of \( L_{\text{spatial}} \) and \( L_{\text{semantic}} \).

The details of our DA-Net are shown in Algorithm 1.

**IV. EXPERIMENTS**

We evaluate our model for the TextVQA task on two challenging datasets, including TextVQA [17] and ST-VQA [2].
We also evaluate our model on the TextCaps dataset [60] for text-based image captioning tasks. Experimental results show that our model achieves superior performance on both TextVQA and TextCaps tasks. All the performances of our model are public on the challenge website of TextVQA, STVQA and TextCaps.

### A. Implementation Details

Following M4C [21], our input includes three parts: question tokens, object tokens, and OCR tokens. For question features, we use three layers of BERT [52] to extract features from question tokens. The BERT layers are finetuned during training. For object features, we use ResNet-152 [61] based Faster-RCNN model [55] to extract object regions and their bounding boxes. For OCR features, we use the Faster-RCNN OCR system as our OCR recognition backbone.

For OCR-CC (1.4M) for visual-linguistic pretraining.

Recent works including TAP [25] and LOGOs [56] apply large-scale pretraining datasets on TextVQA. Here we define the TextVQA pre-trained model as the model using large-scale multi-modality pretraining datasets including IDL (64M) and OCR-CC (1.4M) for visual-linguistic pretraining.

### C. Comparison With The State-of-The-Art

1) **TextVQA**: The TextVQA dataset [17] contains 28,408 images from the Open Images dataset [62], with questions asking about text in the image. Each question in the TextVQA dataset has 10 free-response answers. Following the M4C [21], we multiply the answer as well as its bounding box location.

2) **STVQA and TextCaps**: Our DA-Net builds on the SSbaseline structure. Compared with previous state-of-the-art model BOV [5], we use the SBD-Trans OCR system as our OCR recognition backbone. Our DA-Net builds on the SSbaseline structure. Compared

| Data & Time       | #     | Method           | Data & Time       | Method           | Acc. on val | Acc. on test | Acc. on 3D Subset | Acc. on val | Acc. on test |
|-------------------|-------|------------------|-------------------|------------------|-------------|--------------|-------------------|-------------|--------------|
| 1.4M Data 520 GPU hours | Pretrained Models |              |                      | OCR system      |             |              |                   |             |              |
| 1                 | LOGOs [56] | Multiple-OCR     | 1.4M Data 520 GPU hours | LOGOs [56] | 51.5        | 51.1        | -                 | ✓           | 51.53        | 51.08        |
| 2                 | TAP [25]  | MicroOCR         | 1.4M Data 520 GPU hours | TAP [25]  | 54.7        | 53.9        | -                 | ✓           | 50.57        | 50.71        |
| 28K Data 26 GPU hours | Non-Pretrained Models |              |                      | OCR system      |             |              |                   |             |              |
| 3                 | M4C [21]  | Rosetta-en       | 28K Data 26 GPU hours | M4C [21]  | 39.4        | 39.0        | 42.3               | ✓           | 40.60        | 40.50        |
| 4                 | LaAP-Net [51] | Rosetta-en     | 28K Data 26 GPU hours | LaAP-Net [51] | 40.7        | 40.5        | -                 | ✓           | 41.00        | 41.40        |
| 5                 | SMA [7]   | Rosetta-en       | 28K Data 26 GPU hours | SMA [7]   | 40.0        | 40.6        | -                 | ✓           | 44.60        | 45.50        |
| 6                 | LATR-Base [57]  | Rosetta-en     | 28K Data 26 GPU hours | LATR-Base [57]  | 44.1        | -           | -                 | -           | -            | -            |
| 7                 | TAP [25]  | Rosetta-en       | 28K Data 26 GPU hours | TAP [25]  | 44.1        | -           | -                 | -           | -            | -            |
| 8                 | CRN [58]  | Rosetta-en       | 28K Data 26 GPU hours | CRN [58]  | 40.4        | 41.0        | -                 | -           | -            | -            |
| 9                 | PAT [59]  | Rosetta-en       | 28K Data 26 GPU hours | PAT [59]  | 42.8        | 43.4        | -                 | -           | -            | -            |
| 10                | SA-M4C [8] | Google-OCR       | 28K Data 26 GPU hours | SA-M4C [8] | 45.4        | 44.6        | 45.5               | ✓           | 45.40        | 44.60        |
| 11                | SSbaseline [6] | SBD-Trans      | 28K Data 26 GPU hours | SSbaseline [6] | 43.9        | 44.7        | 43.5               | ✓           | 45.53        | 45.66        |
| 12                | BOV [5]   | SBD-Trans        | 28K Data 26 GPU hours | BOV [5]   | 44.8        | 45.6        | -                 | ✓           | 46.24        | 46.96        |
| 13                | DA-Net (Ours) | Rosetta-en   | 28K Data 26 GPU hours | DA-Net (Ours) | 46.8        | 44.3        | 45.7               | -           | -            | -            |
| 14                | DA-Net (Ours) | SBD-Trans      | 28K Data 26 GPU hours | DA-Net (Ours) | 47.2        | 46.6        | 49.2               | ✓           | 47.12        | 47.11        |

**TABLE I**

**Acc(%) on TextVQA Dataset. In This Paper, the Red Colored Numbers Denote the Best Results Across All Non-Pretrained Approaches in Table. The Blue Colored Numbers Denote the Second-Best Results**

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE II
ACC(%) AND ANLS(%) ON STVQA DATASET. HIGHER IS BETTER IN ALL COLUMNS. IN THIS PAPER, THE RED COLORED NUMBERS DENOTE THE BEST RESULTS ACROSS ALL NON-PRETRAINED APPROACHES IN TABLE. THE BLUE COLORED NUMBERS DENOTE THE SECOND-BEST RESULTS

| Data & Time  | #  | Method     | OCR system   | Acc. (%) on val | ANLS (%) on val | ANLS (%) on test | Acc. on 3D Subset |
|-------------|----|------------|--------------|-----------------|------------------|-------------------|------------------|
| 1.4M Data 520 GPU hours | 1 | LOGOs [56] | Multiple-OCR | 48.6            | 58.1             | 57.9              | -                |
|             | 2 | TAP [25]   | MicroOCR     | 50.8            | 59.8             | 59.7              | -                |
| 22K Data 26 GPU hours | 3 | MM-GNN [35] | Rosetta-en   | 16.0            | -                | 20.7              | 17.1             |
|             | 4 | M4C [21]   | Rosetta-en   | 38.1            | 47.2             | 46.2              | 41.2             |
|             | 5 | LaAP-Net [51] | Rosetta-en | 39.7            | 49.7             | 48.5              | -                |
|             | 6 | SMA [7]    | Rosetta-en   | -               | -                | 46.6              | -                |
|             | 7 | SA-M4C [8] | Google-OCR   | 42.2            | 51.2             | 49.5              | 42.4             |
|             | 8 | CRN [58]   | Rosetta-en   | -               | -                | 48.3              | -                |
|             | 9 | SSBaseline [6] | SBD-Trans | 40.0            | 50.8             | 49.7              | 40.9             |
|             | 10| BOV [5]    | SBD-Trans    | 40.2            | 50.0             | 47.2              | -                |
|             | 11| DA-Net (Ours) | SBD-Trans | 48.5            | 57.5             | 53.1              | 50.0             |

with SSBaseline, our model outperforms 2.4% under the TextVQA Val split and outperforms 1.9% under the Test split. Compared with the previous state-of-the-art model BOV, our model outperforms it by 1.5% under the TextVQA Val split and 1.0% under the Test split. We also use a classic OCR system, Rosetta-en, as the OCR backbone. In Table I, our DA-Net outperforms others under the Rosetta-en backbone.

Following previous TextVQA models [5], [6], [7], we also train our model with additional ST-VQA training data. Our DA-Net also outperforms SA-M4C, SMA, SSBaseline, and BOV, setting a new state-of-the-art result of 47.12% validation accuracy and 47.11% test accuracy on TextVQA dataset.

2) ST-VQA: The ST-VQA dataset [2] is another recently proposed dataset for the TextVQA task. It contains 18,921 training and 2,971 test images sourced from COCO, ICDAR, IIIT, ImageNet, Visual Genome, and VizWiz datasets. Following M4C, we report results on the hardest Open Dictionary (Task-3) as it matches the TextVQA setting where no answer candidates are provided at test time.

Table II shows the performance of our model and baselines on the ST-VQA dataset. We use SBD-trans as our OCR detection system. Following prior works, we show our model’s VQA accuracy and ANLS both on the validation set and only the ANLS metric on the test set. On the validation set, our model achieves an accuracy of 48.5% and an ANLS of 57.5%, which has 8.3% and 7.5% absolutely higher than BOV. On the test set, our model achieves an ANLS of 53.0%, 3.3% better than SSBaseline and 5.8% better than BOV. Overall, our model achieves state-of-the-art performance on test split.

3) Spatial Subsets of TextVQA and STVQA: This experiment is performed on spatial subsets to further explore the TextVQA models’ ability to address spatial-related problems. Inspired by SA-M4C [8], we extract all the questions with “on, top, under” in the valid split of the TextVQA and STVQA datasets and generate the 3D subsets for both datasets. However, the SA-M4C subset [8] considers spatial reasoning as well as OCR matching and contains 409 questions, while our 3D subset only considers 3D spatial reasoning and contains 1766 questions. The statistical summary of our 3D spatial subset is shown in Table III. We re-evaluate our model and baselines on the new 3D subset. As shown in the “Acc. on 3D Subset” column of Table I and Table II, DA-Net achieves the best results under both 3D subsets, surpassing the second by 3.8% and 7.6% on the TextVQA and STVQA subset.

4) Comparison With Pretrained Models: As shown in Table I and Table II, our DA-Net (a non-pretrained method) achieves competitive performance compared to the pretrained models, for example, 48.5% vs. 48.6% by LOGOs [58] and 50.8% by TAP [27] in accuracy, although these pretrained models spend a quite large scale of training data and GPU hours. Specifically, compared to a non-pretrained method (i.e., our method), the scale of the training data for the pretrained models in the TextVQA task is about 70 times, and the GPU hours are about 20 times. In other words, our method is much more efficient and cheaper compared to the pretrained models.

D. Ablation Study
To further verify the performance of every module, we make ablation studies on TextVQA valid split and its 3D subset. Table V shows every module’s influence on the total model. Fig 11 shows the model’s performance under various degrees of noise interference.

1) Effect of Depth Information: To verify the direct improvement of model performance by adding depth information, we remove our depth-aware attention calibration module and the relation prediction task, only using depth information
to supplement the 2D bounding box. As shown in Table V line 2, our baseline model achieves 43.6% on the validation split. Compared with the baseline, our model drops 0.3% because the model cannot comprehend the meaning of depth information without supervision. When bringing the depth information directly into the model, the depth information may be treated as noise information, and thus disturbs the model’s spatial reasoning process.

2) Effect of Depth-Aware Attention Calibration Module: To verify the effectiveness of our depth-aware attention calibration (DAC) module, we remove the relation prediction task and apply the depth-aware attention calibration module on the OCR-visual part (Visual), OCR-semantic part (Semantic), and on both parts (Both). Our model achieves 45.1% when adding depth-aware attention calibration on the OCR-visual part and achieves 44.7% when adding calibration on the OCR-semantic part. Compared with the baseline, our model has a 0.8-1.2% improvement on the validation split. When adding the attention calibration module on both OCR parts, our full depth-aware attention calibration module achieves 46.4% with a 2.5% improvement. The performance of the DAC module is more obvious on the 3D subset of the validation split. This proves the accuracy increase mainly comes from calibrating the OCR attention score according to the spatial relation of the critical object in spatial-related questions. As analyzed in the following subsection, our depth-aware attention calibration module can transfer the attention weight from key objects to their spatial corresponding text tokens. The DAC module guides the model toward the potential answers and restraints from distractor tokens.

3) Effect of Relation Prediction Module: To verify the effectiveness of our relation prediction auxiliary loss, we remove the depth-aware attention calibration module from the whole model. As shown in Table V, our model with only relation prediction modification achieves 45.0% ANLS on validation split, outperforming the baseline by 1.1%. In the 3D subset validation, our model with relation prediction modification surpasses the baseline at 3.6%. The results show that adding the relation prediction task helps the model to understand the 3D geometric information. Inspired by Adabins [39], we also change the spatial loss into bin classification form. Results show that bin classification is worse than the regression form.

We also add experimental results and related descriptions by considering the relationship between OCRs or between OCRs and objects in the relation prediction module. The additional results are shown in Table IV. $L_{\text{Spatial}}^{\text{Obj-Obj}}$ considers the relationship between objects. $L_{\text{Spatial}}^{\text{OCR-OCR}}$ considers the relationship between objects and OCRs. $L_{\text{Spatial}}^{\text{OCR-Obj}}$ considers the relationship between OCRs. Compared with our original relation prediction module which only considers the relationship between objects, the model with $L_{\text{Spatial}}^{\text{Obj-Obj}}$ and $L_{\text{Spatial}}^{\text{OCR-OCR}}$ has better performance, representing that considering Object-OCR spatial relationships benefits the reasoning performance. However, the questions in the TextVQA task rarely involve spatial relationships between OCRs. The addition of $L_{\text{Spatial}}^{\text{OCR-OCR}}$ results in slight performance degradation.

4) Ablation for Surrounding Rate: We evaluate the size of the surrounding area range $\epsilon \in [0, 1]$ setting for the TextVQA dataset validation split in Fig. 7(a). We evaluate the model without the surrounding area ($\epsilon = 0$), only using the center pixel as the depth information. Experiments show that the model without the surrounding area suffers severe performance drop ($\epsilon = 0$). The model with 0.1 surrounding area performs the best, while other surrounding area options have slight degradation in performance.

To explain the performance change with different surrounding areas, we visualize the standard deviation of the depth of all objects in randomly selected 200 images from the TextVQA dataset. Fig. 7(b) shows that when we use the depth of the center pixel as the depth of the total object, the standard deviation substantially increases due to the estimation noise. By using the surrounding area’s average depth as the depth of the total object, the standard deviation can be stable. Comparing the standard deviation of the surrounding area with the different rates in Fig. 7(b), we find that with the increase of the surrounding area rate, the standard deviation decreases, which means the depths between different objects become similar. This makes it difficult for the model to understand

---

**Table IV**

The Effect of the Relation Prediction Module in Our DA-Net

| $L_{\text{Spatial}}^{\text{Obj-Obj}}$ | $L_{\text{Spatial}}^{\text{Obj-OCR}}$ | $L_{\text{Spatial}}^{\text{OCR-OCR}}$ | Valid Acc. |
|-----------------------------------|-----------------------------------|-----------------------------------|-----------|
| √                                 | √                                 | √                                 | 47.2      |
| √                                 | √                                 |                                   | 46.8      |
|                                   | √                                 |                                   | 46.0      |
|                                   |                                   |                                   | 47.7      |
|                                   |                                   |                                   | 47.0      |

---
Fig. 8. Visualization of object’s and OCR’s spatial representations learned by the ssbaseline (without relation prediction module (RPM)) and our DA-Net (with relation prediction module). Comparing the cluster map, our model can successfully split both foreground and background features while the baseline cannot. With RPM the model can learn a clear dividing line between foreground and background features. The RPM promotes the model’s spatial understanding.

Fig. 9. Visualization of OCR attention scores for our model with and without the Depth-aware Attention Calibration (DAC). Attention Calibration represents the visualization of additional scores after the calibration of our module. Our DAC successfully calibrates our model’s OCR attention and focuses on the correct OCR token in the two cases. Final Attention is the visualization of all OCR’s attention scores after the calibration of our module. OCR Attention Change shows the OCR attention score change with and without the depth-aware attention calibration. Our DAC can make the proper calibration, improving the score of correct OCR (Red) and reducing the distractor OCR score (Blue).

| Method       | Depth Info | Spatial Relation | DAC           | Spatial Loss | Val Acc. | Acc. on Subset |
|--------------|------------|------------------|---------------|--------------|----------|----------------|
| SSbaseline   | ✗          | 2D               |               |              | 43.9     | 43.5           |
| SSbaseline   | ✓          | 3D               |               |              | 43.6(−0.3) | 43.7(+0.2)     |
| Ablation     | ✗          | 2D+Overlap       | Visual        | 44.8(+0.9)   | 45.4(+1.9) |
| Ablation     | ✓          | 3D               | Semantic      | 45.1(+1.2)   | 47.6(+4.1) |
| Ablation     | ✓          | 3D               | Both          | 44.7(+0.8)   | 46.4(+2.9) |
| Ablation     | ✓          | 3D               | Bin           | 46.4(+2.5)   | 48.2(+4.7) |
| Ablation     | ✓          | 3D               | Regression    | 44.1(+0.2)   | 44.9(+1.4) |
| DA-Net       | ✓          | 3D               | Full          | 45.0(+1.3)   | 47.1(+3.6) |

TABLE V

Ablation Study for Our Depth-Aware Attention Calibration Module (DAC) and Relation Prediction Module (Spatial Loss)

the spatial relationships between objects. We also visualize the difference surrounding area rate in Fig. 7(c). The green area is the background area. Comparing the $\varepsilon = 0.1$, with the increase of the rate, the surrounding area contains more background area, which makes the depth prediction of the object closer to the depth of the background.
E. Qualitative Analysis

In this part, we cluster the object and OCR features in high dimensions to show the model’s 3D understanding improvement after the addition of the relation prediction task. We visualize heat maps of objects and OCR attention scores to demonstrate the effectiveness of our depth-aware attention calibration module. And we select representative cases to show our model’s ability to address spatial-related questions. We use the baseline model: SSbaseline for comparison.

1) Relation Prediction Module: To demonstrate that our relation prediction module can strengthen the model’s understanding of the 3D spatial relationship between objects, we visualize the spatial clustering of objects and OCRs with and without the relation prediction module in Fig. 8. We choose images with obvious foreground and background split as our visualization cases. Green represents objects and OCRs in background and blue represents objects and OCRs in foreground. To ensure that the clustering is only related to spatial information instead of semantic information, we only visualize objects with the same class: person. We use PCA as the visualization tool.

As shown in Fig. 8, without the relation prediction module, the baseline cannot distinguish the spatial relation between foreground and background (the overlap between green and blue area). After adding the relation prediction module, our model successfully splits the features from the foreground and background. The clustering shows the relation prediction module helps the reasoning model to establish an implicit spatial structure. With RPN, our model has a better 3D understanding of the TextVQA task.
TABLE VI
THE PERFORMANCE ON TEXTCAPS [60]

| Method   | BLEU-4 | METEOR | ROUGE | SPICE | CIDEr |
|----------|--------|--------|-------|-------|-------|
| CNMT[63] | 19.97  | 20.91  | 44.37 | 13.52 | 93.03 |
| BTD[64]  | 20.10  | 17.80  | 42.90 | 11.70 | 41.90 |
| UT[65]   | 20.02  | 20.89  | 44.41 | 13.74 | 85.64 |
| AoAnet[66] | 20.40 | 18.90  | 42.90 | 13.20 | 42.70 |
| M4C-Cap[21] | 23.30 | 22.00  | 46.20 | 15.53 | 89.60 |
| SSbase[6] | 23.53  | 22.02  | 46.40 | 15.01 | 90.50 |
| DA-Net   | 24.35  | 22.47  | 46.95 | 15.32 | 96.80 |

2) Depth-Aware Attention Calibration (DAC): Our depth-aware attention calibration module transfers the object’s attention score to calibrate OCR’s attention score under the supervision of 3D spatial relations. This procedure enables the model to calibrate the OCR’s attention using critical object information. In Attention Calibration, the second column of Fig. 9, we visualize the OCR’s additional attention scores during the transferring process. To make the picture clearer, we show the heatmap of the top three additional attention scores.

To show the change in attention scores after adding our DAC module, we visualize the OCR’s attention scores before and after transferring process in OCR Attention Change, the fourth column of Fig. 9. We only show the top two OCR tokens with noticeable attention score changes to make the picture clearer. For both two cases, the transferred attention score correctly locates the key OCR tokens (“west” and “virginia”). As the OCR Attention Change column shows, for the first case, the attention score of the correct answer token “West” increases by 0.13 after adding the DAC module. For the second case, we find out that our DAC module can also restrain the attention score for distractor token (“325 C1”), which guides the model to answer the correct answer.

3) Case Study: About 44.1% of the questions in the TextVQA dataset need one or more spatial reasoning (Fig. 3). We present several cases with questions that need spatial reasoning for further analysis. As shown in Fig. 10, our model performs well in challenging cases with the supervision of 3D spatial information and a step-by-step spatial reasoning procedure.

The first case on line 1 and line 2 of Fig. 10 shows that our model can better predict OCR answers with object-text spatial relations, while the baseline fails to generate the entirely correct answer. The second and third cases on line 1 and line 2 of Fig. 10 show that our model can better locate correct objects while the baseline tends to locate the distractors.

However, there are cases in which both our DA-Net and the baseline cannot predict the answer correctly. (i) Images with complex semantic information in OCR tokens. As shown in the first case on line 3, when questions aim at the semantic information in the OCR token, existing TextVQA models tend to predict unrelated OCR in the image as the answer. (ii) Images with false OCR information. As shown in the second and third cases on line 3, when the OCR detection module generates incorrect OCR tokens, our downstream reasoning model cannot predict correct answers.

F. Generalization Analysis
In addition to TextVQA, our model can also be applied to other scene-text tasks including the TextCaps task. This task requires models to generate image descriptions via texts in the context of images. In this paper, we conduct experiments on the TextCaps dataset [60] for text-based image captioning. The TextCaps dataset, with 145k captions for 28k images, is recently proposed for the text-based image captioning task. Following M4C-Captioner [60], we use BLEU-4 [67], METEOR [68], ROUGE [69], SPICE [70] and CIDEr [71] to evaluate the performance of captioning models. All automatic metrics are positively correlated with the generated quality. Table VI shows that DA-Net outperforms others and achieves a new state-of-the-art performance of 96.80 CIDEr-D score and 24.35 BLEU-4 score. The performance demonstrates our model’s generalization ability on other scene-text tasks.

V. CONCLUSION AND FUTURE WORK
In this paper, we study the spatial reasoning problem from human-like spatial reasoning for TextVQA. To this end, we propose the Depth-Aware TextVQA Network (DA-Net) by introducing the 3D depth information into the TextVQA task. We introduce a relation prediction module to enhance the model’s understanding of 3D spatial relations. Besides, we design a depth-aware attention calibration module to readjust the distribution of OCR tokens’ attention scores based on the crucial object. The experiments on two TextVQA datasets demonstrate the effectiveness of our method. Further, we also apply 3D spatial reasoning in our work to other visual reasoning tasks such as text-based image captioning.

VI. APPENDIX
A. Extra Ablation Studies
1) Ablation for Object Number: For every input image, we detect a fixed number of objects from the image. In order to explore the impact of the detected object number, we make experiments about our DA-Net with the fixed detected objects N=36, N=80, and N=100 via bottom-up attention on the TextVQA dataset. As shown in Table VIII, we find that the N=100 setting and the N=80 setting correspond to almost the same performance. But when we use the N=36 setting, the model suffers slight degradation in performance. Thus, we apply N=80 as our final object number, which takes into account both performance and computation complexity.

2) Ablation for Vocabulary Size: We analyze the question type, answer number, and vocabulary size of the TextVQA dataset and the STVQA dataset. We further explored the effect of different vocabulary sizes on performance.

For the TextVQA dataset, the valid split has 3450 different answers. However, different from VQA, the answer in TextVQA contains multiple words. For example, an answer from TextVQA, “Star War”, contains 2 words. Thus, the vocabulary in the TextVQA is much larger than 3450. The actual vocabulary size in the TextVQA train+valid split is 46009, far larger than our 5000 fixed vocabulary size. We find the original Table C is misleading. We update the manuscript using Table IX to replace the original Table C.

We also make ablation experiments with different vocabulary sizes. vocab_3k is the top 3000 words in the valid split. fixed_vocab_5k is the fixed vocabulary we originally use in our model. vocab_7k is the top 7000 words in the valid split.

From the experimental results in Table VII, we find that the performance of the model is similar under the vocab_3k and vocab_5k settings. The number of commonly-used words in TextVQA is 2844 (Appearing more than 10 times in the
Fig. 11. More cases that 2D spatial modeling cannot get correct predictions.

### TABLE VII

| Vocabulary Size | TextVQA | STVQA |
|-----------------|---------|-------|
|                 | vocab_3k | fixed_vocab_5k | vocab_7k | vocab_3k | fixed_vocab_5k | vocab_7k |
| Acc.            | 47.0     | 47.2   | 46.5     | 48.6     | 48.5     | 47.7     |

### TABLE VIII

| The fixed detect objects N | N=36 | N=80 | N=100 |
|----------------------------|------|------|-------|
| Accuracy                   | 46.1 | 47.1 | 47.2  |

### TABLE IX

| Dataset  | Ques-Type | Ans-Num | Vocab   |
|----------|-----------|---------|---------|
| TextVQA  | Valid     | 14      | 3450    | 46009   |
|          | 3D-Val    | 12      | 1150    | 43066   |
| STVQA    | Valid     | 12      | 2300    | 15884   |
|          | 3D-Val    | 8       | 868     | 13294   |

TextVQA dataset). The vocab_3k and vocab_5k settings both contain the commonly-used words. Thus, the performance is similar. Also, we find that a large vocabulary causes performance degradation. The performances with vocab_7k in TextVQA and STVQA are lower than in other settings. The increase in the vocabulary size leads to the increase of prediction vector dimension, which makes the prediction harder.

#### B. Extra Visualization

To further illustrate the problem in the 2D spatial reasoning process, we list more cases that 2D spatial modeling that fails to predict the correct answer in Figure 11.

- **The first case:** The question asks the name of the box. From the 2D view, both OCR tokens, “Fosters” and “corona”, overlap the box region. The 2D spatial reasoning model cannot distinguish and predict the wrong answer.
- **The second case:** The question asks the word on the carriage behind the car. From the 2D view, both OCR tokens, “Budget” and “GMC”, overlap the carriage. The 2D spatial reasoning model cannot distinguish and predict the wrong answer.
- **The third case:** The question asks the word on the player’s shirt. From the 2D view, the OCR token “Karmann” is inside the player region, and the OCR token “eddielemonsportingimages.com” has a large area overlapping the player region. The 2D spatial reasoning model cannot distinguish and predict the wrong answer.
- **The fourth case:** The question asks the brand of the flag. From the 2D view, the OCR tokens including “FIN”, “jeep”, and “Engle” overlap the flag. The 2D spatial reasoning model cannot distinguish and predict the wrong answer.

#### C. Robustness to Depth Noise

The depth information predicted by AdaBins [39] contains slight noise. To verify the depth noise influence on our model, we randomly mask the input depth information and Table XI shows the results. Specifically, we replace the original depth information using the mean value of all depth information as the noise. After masking 80% and 40% depth information, the performance drops significantly. However, after masking 20% and 10% depth information, the performance only suffers little volatility and outperforms the SSbaseline model. It indicates that our model is able to suffer slight depth noise.
D. Computational Complexity

We list the computational complexity of our DA-Net and other TextVQA models including SSBaseline, M4C, and BOV in Table F. Specifically, \( L = 20 \) is the length of the question; \( N = 50 \) is the number of the OCR tokens; \( M = 100 \) is the number of the detected objects; \( C = 10 \) is the number of the candidate answers in BOV. We omit all vector dimensions \( D \) for simplicity.

Due to the addition of the depth-aware attention calibration module, our DA-Net’s attention block complexity is slightly larger than other models. Table X also shows the complexity of the MMT encoder and the comparison of the encoder’s total computation complexity (the summary of the Transformer encoder and attention block complexity). Our DA-Net structure’s computational complexity is much smaller than the complexity of M4C and BOV. Our depth-aware attention calibration module and the relation prediction module create a little computation complexity increase.

E. Limitation of DA-Net

The main limitation of DA-Net is that the depth estimation module could produce noisy depth information and is not well designed for TextVQA. Our future work will integrate the depth estimation module into the DA-Net to adapt the TextVQA task in an end-to-end manner, which may further boost the spatial reasoning ability of our model.

ACKNOWLEDGMENT

This work is distributed under the LaTeX Project Public License (LPPL) (http://www.latex-project.org/) version 1.3. A copy of the LPPL, version 1.3, is included in the base LaTeX documentation of all distributions of LaTeX released 2003/12/01 or later. The opinions expressed here are entirely that of the author. No warranty is expressed or implied. User assumes all risk.

REFERENCES

[1] A. Mishra, S. Shekhar, A. K. Singh, and A. Chakraborty, “OCR-VQA: Visual question answering by reading text in images,” in Proc. Int. Conf. Document Anal. Recognit. (ICDAR), Sep. 2019, pp. 947–952.
[2] A. F. Biten et al., “Scene text visual question answering,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 4291–4301.
[3] A. Singh et al., “Towards VQA models that can read,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 8309–8318.
[4] Y. Qiu, Y. Satoh, R. Suzuki, and H. Kataoka, “Incorporating 3D information into visual question answering,” in Proc. Int. Conf. 3D Vis. (3DV), Sep. 2019, pp. 756–765.
[5] G. Zeng, Y. Zhang, Y. Zhou, and X. Yang, “Beyond OCR + VQA: Involving OCR into the flow for robust and accurate TextVQA,” in Proc. 29th ACM Int. Conf. Multimedia, Oct. 2021, pp. 376–385.
[6] Q. Zhu, C. Gao, P. Wang, and Q. Wu, “Simple is not easy: A simple strong baseline for TextVQA and TextCaps,” 2020, arXiv:2012.05153.
[7] C. Gao et al., “Structured multimodal attentions for TextVQA,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, pp. 9603–9614, Dec. 2022.
[8] Y. Kant et al., “Spatially aware multimodal transformers for TextVQA,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2020, pp. 715–732.
[9] C. Helwe, C. Clavel, and F. M. Suchanek, “Reasoning with transformer-based models: Deep learning, but shallow reasoning,” in Proc. 3rd Conf. Automated Knowl. Base Construction, 2021, pp. 1–11.
[10] P. Banerjee, T. Gokhale, Y. Yang, and C. Baral, “Weakly supervised relative spatial reasoning for visual question answering,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 1888–1898.
[11] L. Vieu, “Spatial representation and reasoning in artificial intelligence,” in Spatial Temporal Reasoning, Springer, 1997, pp. 5–41.
[12] A. Chang, M. Savva, and C. D. Manning, “Learning spatial knowledge for text to 3D scene generation,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2014, pp. 2028–2038.
[13] Y. Zhou et al., “SP-VIT: Learning 2D spatial priors for vision transformers,” 2022, arXiv:2206.07662.
[14] H. Wu et al., “CVT: Introducing convolutions to vision transformers,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 22–31.
[15] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, “Incorporating convolution designs into visual transformers,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 559–568.
[16] L. Yuan et al., “Tokens-to-Token VIT: Training vision transformers from scratch on ImageNet,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 538–547.
[17] A. Singh et al., “Towards VQA models that can read,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 8309–8318.
[18] H. Li, X. Li, B. Karimi, J. Chen, and M. Sun, “Joint learning of object graph and relation graph for visual question answering,” 2022, arXiv:2205.04188.
[68] M. Denkowski and A. Lavie, “Meteor universal: Language specific translation evaluation for any target language,” in *Proc. 9th Workshop Stat. Mach. Transl.*, 2014, pp. 376–380.

[69] L. C. Rouge, “A package for automatic evaluation of summaries,” in *Proc. Workshop Text Summarization ACL*, 2004, pp. 1–10.

[70] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: Semantic propositional image caption evaluation,” in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 382–398.

[71] R. Vedantam, C. L. Zitnick, and D. Parikh, “CIDEr: Consensus-based image description evaluation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.

Hao Li received the bachelor’s degree from the School of Computer Science, Peking University, Beijing, China, in 2021, where he is currently pursuing the Ph.D. degree with the School of Computer Science. His research interests include cross-modality learning and its downstream task, such as video-text retrieval, image caption, and visual question answering.

Jinfa Huang received the bachelor’s degree (Hons.) from the University of Electronic Science and Technology of China, Chengdu, in 2020. He is currently pursuing the master’s degree with Peking University, Beijing, China. His research interests encompass cross-modality learning and language, with a focus on visual captioning, visual question answering, and video-text retrieval.

Peng Jin (Graduate Student Member, IEEE) received the B.E. degree from the School of Electronic Information, Tsinghua University, China, in 2021. He is currently pursuing the master’s degree with Peking University, Beijing, China. His research interests include cross-modality learning and its downstream task, such as video-text retrieval, image caption, and visual question answering.

Guoli Song received the B.S. degree in mathematics and applied mathematics and the M.S. degree in operational research and cybernetics from Zhengzhou University in 2009 and 2012, respectively, and the Ph.D. degree in computer engineering from the University of Chinese Academy of Sciences in 2018. She is currently a Junior Research Scientist with the Peng Cheng Laboratory. Her research interests include cross-modal learning, computer vision, and medical image analysis.

Qi Wu (Member, IEEE) received the bachelor’s degree in mathematical sciences from China Jiliang University, the master’s degree in computer science, and the Ph.D. degree in computer vision from the University of Bath, U.K. He was a Postdoctoral Researcher with the University of Adelaide in 2012, 2015, and 2017, respectively. He is also an ARC DECRA Fellow from 2019 to 2021. He is currently an Associate Professor with The University of Adelaide. His research interests include cross-depiction object detection and classification, attributes learning, neural networks, and image captioning.

Jie Chen (Member, IEEE) received the M.Sc. and Ph.D. degrees from the Harbin Institute of Technology, China, in 2002 and 2007, respectively. He joined as a Faculty Member with the Shenzhen Graduate School, Peking University, in 2019, where he is currently an Associate Professor with the School of Electronic and Computer Engineering. Since 2018, he has been working with the Peng Cheng Laboratory, China. From 2007 to 2018, he worked as a Senior Researcher with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. In 2012 and 2015, he visited the Computer Vision Laboratory, University of Maryland, and the School of Electrical and Computer Engineering, Duke University, respectively. His research interests include deep learning, computer vision, large language models, and AI4Science. He was the Co-Chair of International Workshops at ACCV, ACM MM, CVPR, ICCV, and ECCV. He was a Guest Editor of special issues of *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IJCV*, and *Neurocomputing*. He has been selected as a Finalist of the 2022 Gordon Bell Special Prize for HPC-Based COVID-19 Research. He is an Associate Editor of the *Visual Computer*. 