From Easy to Hard: Learning Language-Guided Curriculum for Visual Question Answering on Remote Sensing Data

Zhenghang Yuan, *Student Member, IEEE*, Lichao Mou, Qi Wang, *Senior Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

Abstract—Visual question answering (VQA) for remote sensing scene has great potential in intelligent human–computer interaction system. Although VQA in computer vision has been widely researched, VQA for remote sensing data (RSVQA) is still in its infancy. There are two characteristics that need to be specially considered for the RSVQA task: 1) no object annotations are available in the RSVQA datasets, which makes it difficult for models to exploit informative region representation and 2) there are questions with clearly different difficulty levels for each image in the RSVQA task. Directly training a model with questions in a random order may confuse the model and limit the performance. To address these two problems, in this article, a multi-level visual feature learning method is proposed to jointly extract language-guided holistic and regional image features. Besides, a self-paced curriculum learning (SPCL)-based VQA model is developed to train networks with samples in an easy-to-hard way. To be more specific, a language-guided SPCL method with a soft weighting strategy is explored in this work. The proposed model is evaluated on three public datasets, and extensive experimental results show that the proposed RSVQA framework can achieve promising performance. Code will be available at https://gitlab.lrz.de/ai4eo/reasoning/VQA-easy2hard.

I. INTRODUCTION

IMAGES from spaceborne and airborne platforms usually cover large-scale geographical areas and provide important data bases for many Earth observation (EO) applications [1]. With the development of EO technology, there have been an increasing number of works on remote sensing image analysis, such as land use classification [2], [3], object detection [4], [5], road extraction [6], and change detection [7]–[9]. However, due to the specialized nature of remote sensing tasks, the ability to carry out such tasks is limited to experts in the related fields. The obtained semantic information from some remote sensing tasks is not intuitive to common users, which makes it difficult to deliver the image information to users in domain-specific applications.

Fortunately, novel tasks such as image captioning [10], [11] and visual question answering (VQA) [12], [13] have recently been explored for visual data. These tasks take both natural language and imagery as inputs and output easy-to-understand text in natural language. Among them, VQA has become a hot research topic in artificial intelligence community [14]. Given an input image and a natural language question, VQA aims to generate a textual answer to the question based on image content [15]. It is an interdisciplinary research area between computer vision and natural language processing [16]. It is also a challenging task that requires a model to jointly learn multi-modal representation from both imagery and language data. More specifically, a VQA model needs to learn visual representation to understand the input image and effective features for natural language to gain an answer conditioned on image content [17].

As for remote sensing data, VQA enables end-users to better understand a complicated remote sensing image and has great potential in human–computer interaction applications [18]. A pioneer work can be found in [19], where the authors created two datasets and proposed a baseline model of VQA for remote sensing data (RSVQA). Although VQA in computer vision has been widely studied [20]–[23], VQA for remote sensing imagery is still in its infancy. Due to different image characteristics, VQA methods for natural images may not
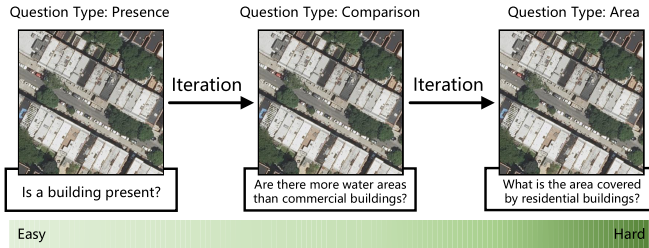
Fig. 1. Motivation of the proposed method: learning features from easy samples to hard ones.

work well on remotely sensed images. Specifically, two main challenges for the RSVQA task are summarized as follows.

1) No object annotations available in the RSVQA datasets. In computer vision, VQA models are able to make use of the existing object annotations to learn features tailored to objects, which helps a lot to improve performance [24], [25]. In contrast, there are no object annotations available in the RSVQA datasets, making it difficult for the RSVQA models to take advantage of informative region information.

2) Questions of different difficulty levels. Questions about each remote sensing image have significantly different difficulty levels. Fig. 1 shows several examples. Usually, upon learning a VQA model, training image–question–answer triplets get shuffled. By doing so, the model can see different samples in no particular order and learn the task evenly without getting stuck in local optima. However, this may confuse the model and hence affect the final performance, as easy and difficult questions are in the same batch.

Aiming at the above-mentioned two challenges, our motivations are explained from two aspects. First, both holistic and region features should be well-exploited to enhance visual representation for RSVQA. Though the holistic feature provides global information of the input image, it may neglect some important details, whereas the region feature can provide more detailed semantic information, which is critical for answering complicated questions. Moreover, due to the fact that remote sensing images usually contain objects of various scales, using region representation is also helpful for addressing the scale-variation problem. To harness both the features, we propose a multi-level visual feature learning method. Specifically, the language-guided holistic image feature and the region feature are jointly learned to improve the performance of RSVQA models.

Second, the model should be trained in ascending order of learning difficulty. We humans tend to learn from easy to hard. Inspired by the human learning process, self-paced curriculum learning (SPCL) is explored in this work and shows promising results. It considers question attributes and model feedback to dynamically adjust the question sequence for model training in ascending order of difficulty [26], [27], namely, from easy samples to hard ones. Albeit successful in many problems, SPCL for the RSVQA task still remains underexplored.

To sum up, the main contributions of this work can be summarized as follows.

1) A multi-level visual feature learning method is proposed to jointly exploit both holistic and region features. Specifically, a cross-modal global attention (CGA) module is devised to learn the language-guided holistic image feature, and a cross-modal spatial transformer (CST) module is developed to learn the question-related region feature.

2) The proposed CST module applies affine transformation to visual features to automatically crop informative regions without object annotations. Moreover, the language feature is also used as guidance to generate multiple spatial transformation parameters for obtaining richer region features.

3) A language-guided SPCL method with a soft weighting strategy is devised for RSVQA. It takes question length and type as prior knowledge and dynamically adjusts question sequence to enable a more effective training process: learning with easy questions and then with hard ones.

The rest of this article is organized as follows. Related works about VQA for both natural images and remote sensing images are introduced in Section II. The methodology is described in Section III. Section IV presents the experimental results and discussion. Finally, this article is concluded in Section V.

II. RELATED WORK

Multi-modal feature learning [28]–[30] plays an important role in both remote sensing and computer vision tasks. For a typical VQA framework, learning multi-modal representation is also one of the core components. Malinowski and Fritz [31] combined semantic segmentation of scenes and symbolic reasoning over questions to learn multi-modal features. With the development of deep learning, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are usually used for visual and language feature encoding and have become mainstream feature learning methods [17], [32], [33]. Antol *et al.* [15] introduced the task of free-form and open-ended VQA and used VGGNet and LSTM to extract multi-modal features.

Besides, visual attention mechanism is also widely used in VQA tasks to make the model focus on important pixels. Chen *et al.* [34] proposed a language-guided attention method that projects question embeddings into a visual space and to learn multi-modal features. Fukui *et al.* [35], Kim *et al.* [36], and Ben-Younes *et al.* [37] designed different multimodal bilinear pooling methods to integrate visual features with language features. Yu *et al.* [38] further reduced the co-attention method into question self-attention and question-conditioned attention for learning better visual features. However, softly attended visual features are still holistic representations of the image. The detailed region information is neglected by these methods, which is critical for alleviating the scale-variation problem in the RSVQA task.

To leverage the object-level semantic information of the input image, patch-based and object-based feature learning methods are proposed to extract more representative local features. For VQA models in computer vision, object detectors
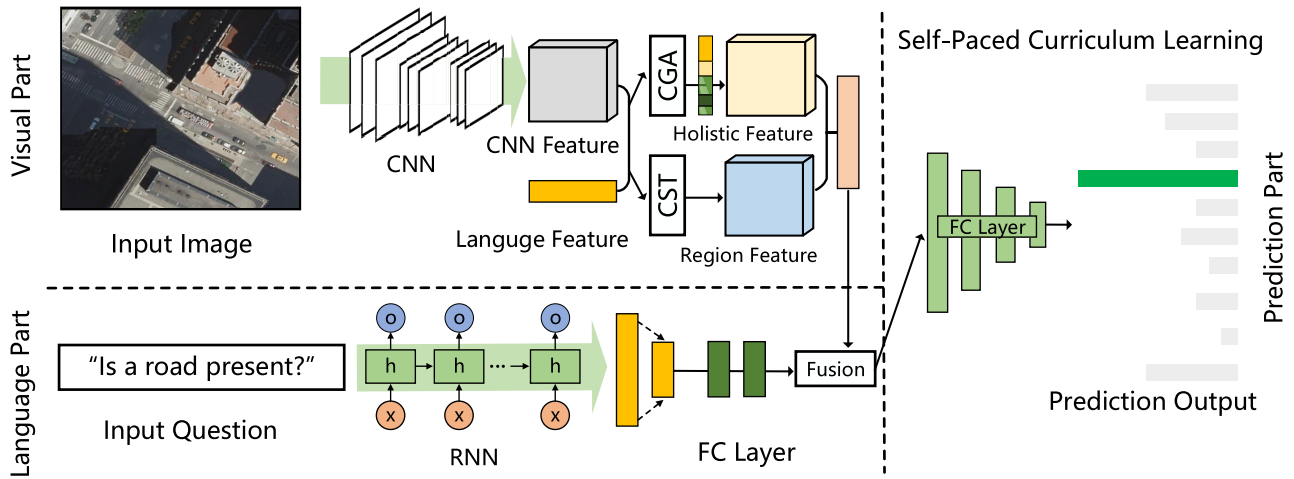
Fig. 2. Main architecture of the proposed VQA method. 1) First, multi-modal features are extracted from the two types of inputs, including visual features from the given image and language features from the question; 2) then, visual features and language features are somehow fused to get the multi-modal representation; and 3) finally, the answer is predicted via a classifier.

are usually used to represent the image as a collection of bounding boxes [39]–[41]. Recently, vanilla grid convolutional feature maps [42] are also proven to be effective for visual feature learning in VQA and image captioning tasks. However, these methods all rely on bounding box annotations, which are not available in the RSVQA datasets.

Compared with VQA for natural imagery, the research for RSVQA is still in its early stage. The first work for VQA on remote sensing data was introduced by Lobry *et al.* [19], where a template-based automatic method was designed to build two remote-sensing-oriented datasets. The image–question–answer triplets of the two datasets were constructed via the information from OpenStreetMap and pre-defined templates. They used a CNN to extract visual features and an RNN to learn language features. After point-wise fusion of multimodal features, an answer is predicted by a classification task. This work paves the way for RSVQA by providing datasets and a baseline method for further research. However, due to unique characteristics of remote sensing imagery, more specific feature learning algorithms need to be investigated and explored for this task.

III. METHODOLOGY

As mentioned above, this work focuses on two problems. On one hand, how can we adaptively exploit both holistic and regional visual features for answering different types of questions. On the other hand, how can we more effectively train a model with questions of different difficulty levels. The whole architecture of the proposed RSVQA framework is shown in Fig. 2. It consists of two parts: 1) CGA and CST modules that intend to learn multi-level visual features with multiple spatial contexts and 2) SPCL-based network training, which aims to train a model in an easy-to-hard way. Following prior works [19], [43], we formulate predicting the answer conditioned on the input image and question as a classification task instead of a sentence generation task. Specifically, the final answer can be selected from the class with the highest probability. In Sections III-A and III-B, the two parts of our RSVQA framework will be described in detail.

A. Multi-Level Visual Feature Learning With Multiple Spatial Contexts

Compared with natural images, remote sensing data usually contain much richer content due to top–down views, which enables people to ask various types of questions for the same input image. The existing RSVQA datasets in [19] contain the following five types of questions, and an example is given for each type.

1) *Rural/Urban:* "Is it a rural or an urban area?" The answer to this question can be deemed as a typical binary classification task.
2) *Presence:* "Is a road present?" To answer presence-related questions, a VQA model needs to predict whether there exist specific objects.
3) *Comparison:* "Are there more roads than residential buildings?" We need to compare the areas or objects involved.
4) *Area:* "What is the area covered by residential areas?" The model needs to seek out the target according to the question.
5) *Count:* "What is the amount of small buildings?" To answer count-related questions, the numbers of specific objects need to be predicted.

As can be seen, different image regions and spatial contexts need to be used to answer various types of questions. Thus, we propose to combine holistic (global) and region (local) features to learn multi-level visual representation with multiple spatial contexts. Specifically, CGA is designed to extract global visual features with the guidance of language. CGA uses language features as the guidance to generate global attention maps on the whole image. Meanwhile, the CST module is proposed to extract region features in an adaptive way. CST learns to spatially transform feature maps to be of different scales and poses, and the transformed features can be used to enhance visual representations. Details of the two proposed modules are illustrated in Fig. 3 and are introduced in Section III-A.
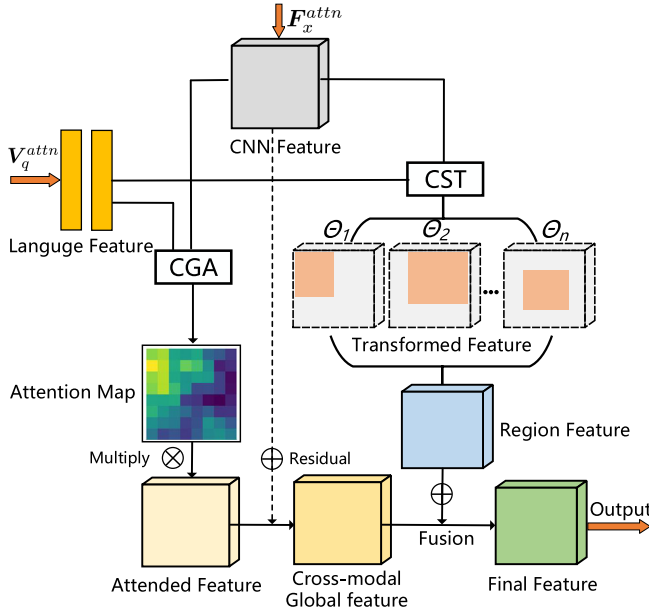
Fig. 3. Illustration of the proposed CGA and CST visual feature learning modules. Using the proposed two modules, cross-modal global and spatially transformed visual features can be learned jointly.

1) CGA Module: Attention mechanism has shown its effectiveness in many computer vision tasks [28], [44], [45]. This is mainly because focusing on some important regions of an image can improve the discriminability of visual features. For object recognition or semantic segmentation tasks, only one modality, i.e., vision, is input to the model, and self-attention is used. However, there are two modalities in our case. Therefore, we propose a cross-modal global attention, i.e., CGA, which exploits the language feature as guidance to generate global attention maps on all the locations.

Formally, let x be the input image, and the corresponding question is denoted as q. For visual and language modalities, CNNs and RNNs are commonly used as feature encoders. Correspondingly, visual feature $F_x \in \mathbb{R}^{N \times C \times H \times W}$ and language feature vector $v_q \in \mathbb{R}^{N \times L}$ can be obtained with the networks. N is the batch size, C is the number of channels, and H and W represent the height and width of the input image, respectively. L denotes the dimension of the language feature vector. The typical, single-modality attention mechanism first encodes the visual feature F_x into three independent features: the query Q, key K, and value V. The key idea of attention mechanism is to assign weights to the input value V according to a similarity function. Usually, the attention weights are computed by the compatibility between the input query and the corresponding key. Generally, the attention [44] for the single-modality input can be calculated by

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

where d_k is the number of channels of F_x.

However, for RSVQA, we make use of the information of the natural language question to assign weights for visual features. To this end, we devise a multi-modal compatibility function to compute the similarities between language and

visual features at different locations. First, a 1×1 convolution layer and a fully connected (FC) layer are used to transform the visual and language features into $F_x^{\text{attn}} \in \mathbb{R}^{N \times C \times H \times W}$ and $V_q^{\text{attn}} \in \mathbb{R}^{N \times C}$. Then, we expand the dimension of V_q^{attn} to the same dimension as F_x^{attn}. Afterward, the query Q^{attn} can be computed as

$$Q^{\text{attn}} = \text{Conv}_{1 \times 1}\left(\text{ReLU}\left(\frac{F_x^{\text{attn}}}{\|F_x^{\text{attn}}\|_2} + \frac{V_q^{\text{attn}}}{\|V_q^{\text{attn}}\|_2}\right)\right). \qquad (2)$$

Finally, the cross-modal attention can be defined as

$$\text{Attention}(Q^{\text{attn}}, F_x^{\text{attn}}, V) = \text{softmax}\left(\frac{Q^{\text{attn}} F_x^{\text{attn}^T}}{\sqrt{d_k'}}\right)V + F_x \qquad (3)$$

where the value V is computed by $V = \text{Conv}_{1 \times 1}(F_x)$, and d_k' is the channel dimension of F_x^{attn}.

2) CST Module: Attention mechanism can make model focus on informative image features using soft weights. However, it is still a global feature learning method with a fixed spatial context. Compared with global features, multi-level features are more effective. Therefore, we propose a cross-modal spatial transformer, i.e., our CST module, to extract features with adaptive scales and spatial contexts. The spatial transformation in CST includes cropping, translation, and scaling, which can be learned in an end-to-end manner without object annotations. As opposed to the soft attention in CGA module, the CST module can be viewed as a hard attention method for region feature learning.

The spatial transformer in CST is a differentiable transformation module, and it is conditioned on both visual and language features. Specifically, there are three sub-components in the spatial transformer following the work in [46]: localization network, parameterized grid sampling, and differentiable bilinear sampling.

In this work, the transformation parameter T_θ can be defined as

$$T_\theta = \begin{bmatrix} s_1 & 0 & t_x \\ 0 & s_2 & t_y \end{bmatrix} \qquad (4)$$

where s_1, s_2, and t_x, t_y are the parameters controlling the scaling, translation, and cropping transformation, respectively.

To predict the transformation parameter T_θ, we design a cross-modal localization network. In CGA, F_x^{attn} and Q^{attn} are computed for generating cross-modal attention. In the cross-modal localization network, we reuse F_x^{attn} and V_q^{attn} for predicting the transformation parameter T_θ. Since different feature channels focus on different parts of the image [30], [47], we propose to use multiple spatial transformers from different channel groups to extract richer visual features. This can be defined as

$$M_1^{\text{attn}}, M_2^{\text{attn}} = \text{split}\left(\frac{F_x^{\text{attn}}}{\|F_x^{\text{attn}}\|_2} + \frac{V_q^{\text{attn}}}{\|V_q^{\text{attn}}\|_2}\right). \qquad (5)$$

We split the cross-modal feature into M_1^{attn} and M_2^{attn} along the channel dimension evenly. Then, two transformation

parameters are predicted by split features as

$$T_{\theta_1} = \text{FC}\big(\text{ReLU}\big(M_1^{\text{attn}}\big)\big)$$
$$T_{\theta_2} = \text{FC}\big(\text{ReLU}\big(M_2^{\text{attn}}\big)\big) \tag{6}$$

where FC denotes FC layers. T_{θ_1} and T_{θ_2} denote the predicted transformation parameters.

Note that two spatial transformers share a similar differentiable bilinear sampling process, which can be defined by

$$E_i^c = \sum_u^H \sum_v^W F_x^c(u, v)\max(0, 1-|x_i-v|)\max(0, 1-|y_i-u|) \tag{7}$$

where $i \in \{1, 2, \ldots, WH\}$ is the coordinate index, and c is the channel index. Transformed spatial coordinates (x_i, y_i) and the feature coordinates (u, v) of F_x are normalized in the range of $[-1, 1]$. E_i^c denotes the sampled features using the transformation parameter T_{θ_1} or T_{θ_2}. For each spatial transformer, we compute partial derivatives for both features and coordinates. By this means, the whole networks can be trained in an end-to-end manner.

B. Cross-Modal Feature Learning: From Easy to Hard

Traditional stochastic training strategy usually takes the input samples in a random order, while this is opposite to the learning process of human. Curriculum learning (CL) [48], self-paced learning (SPL) [49], and SPCL [26] are proposed as more reasonable training algorithms for machine learning models. The core idea of them is to train a model starting from easy samples and gradually including hard ones. Previous works [27], [50] show that designing proper ranking functions to organize training samples in ascending order of learning difficulty is helpful for improving the model performance. Since there exist questions with different difficulty levels for the same remote sensing image, training an RSVQA model from easy to hard is a more reasonable strategy.

In this section, we propose a language-guided SPCL training method for RSVQA. Generally, SPCL is composed of two parts: SPL and CL. SPL can be reformulated as an optimization problem, and the curriculum is dynamically adjusted according to model feedback during the training phase. The curriculum in CL is determined by prior knowledge, which is a prejudgment about the difficulty level of a specific task.

In this work, the target of SPL is to adjust the sequence of input samples during the training stage. Specifically, SPL uses adaptive weights for each training sample to control the training sequence by an importance sampling strategy. Let $v = [v_1, v_2, \ldots, v_N]$ denote the weight vector for each sample in N training questions. x_i is the ith input image, and q_i is the ith input question. $g(h(x_i), s(x_i), q_i; w)$ represents the whole RSVQA model. Here, $h(x_i)$ denotes the global feature learned by CGA module, and $s(x_i)$ denotes the transformed feature learned by CST module. w represents the learnable network weights in the whole model. Then, SPL is exploited to train the model with samples organized in ascending order of learning difficulty. Based on the learned multi-level features, the SPL

loss can be defined as

$$\min_{w,v} \mathbb{E}(w, v, \lambda) = \sum_{i=1}^N v_i L\big(y_i, g\big(h(x_i), s(x_i), q_i, w\big)\big) + f(v; \lambda) \tag{8}$$

where samples with larger v have larger influence on model training and vice versa. y_i is the ground-truth label. Since we take answer prediction as a classification task, its loss function is a cross-entropy function represented by $L(y_i, g)$. λ can be interpreted as the "age" of the model, which is used to control the learning pace.

Actually, $f(v; \lambda)$ is a self-paced regularizer for controlling the learning process. The vector v is learnable and updated by optimizing the SPL loss function. Basically, elements of v can be hard (0 or 1) or soft (from 0 to 1). In what follows, we study the soft regularizer for SPL to enable a more flexible training of the RSVQA model. Specifically, the soft regularizer can be defined as follows:

$$f = \lambda\left(\frac{1}{2}v^2 - v\right), \quad v \in (0, 1)^N. \tag{9}$$

Given that there are two disjoint blocks of variables, i.e., network weights w and weight vector v, SPL is a biconvex optimization problem. Usually, the alternative convex search algorithm is used to solve it. When network parameters w including all learnable weights in CGA and CST modules are fixed, the global optimum v^* for the regularizer can be computed by

$$v_i^* = \begin{cases} -\dfrac{L}{\lambda} + 1, & \text{if } L\big(y_i, g\big(h(x_i), s(x_i), q_i, w\big)\big) \leq \lambda \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

After vector v is updated, we fix v and optimize network weights w by a stochastic gradient descent (SGD) optimizer. Since easy samples can be quickly fitted with limited iterations, the loss values for easy samples are usually smaller than those for hard ones. So if the loss value L is not larger than λ, the corresponding input question will be taken as an easy sample and trained with high priority. Otherwise, v_i^* is set to 0, and the corresponding question will not be used for training.

As mentioned, λ is the "age" of the model that increases gradually along with the training iteration. In this work, we record the maximum and minimum loss values of epoch $t - 1$, and use them to update λ as follows:

$$\lambda = \big(\max\big(L^{t-1}\big) - \min\big(L^{t-1}\big)\big) \cdot K + \min\big(L^{t-1}\big) \tag{11}$$

where K is used to adjust the value of λ. Specially, we define K as a dynamic changing parameter for controlling the learning pace. The initial value of K is set to 0.5, and it is updated during the training stage by

$$K = 0.5 + \frac{t}{15} \times 0.1. \tag{12}$$

When the value of λ increases, the model includes more difficult questions with larger loss values.

However, SPL does not incorporate prior knowledge in the learning process. In the initial stage, the network weights are still randomly initialized, and the loss values of easy and hard examples may not be accurate to determine the true difficulty order. Thus, incorporating prior knowledge is necessary in our case. Inspired by CL, we design a curriculum, namely, a ranking function to organize questions in an easy-to-hard order at the beginning of training. By combining SPL with CL, the proposed language-guided SPCL method for RSVQA can take advantages of these two learning regimes.

Two factors are considered in this work to design the ranking function in CL: question length and question type. In most cases, longer questions are usually more complicated than shorter ones. In addition, different types of questions also have different difficulty levels. For example, object recognition is usually easier than counting task.

Based on this prior knowledge, the SPCL loss can be defined as

$$\min_{\boldsymbol{w},\boldsymbol{v}} \mathbb{E}(\boldsymbol{w}, \boldsymbol{v}, \lambda, \Psi) = \sum_{i=1}^{N} v_i L\big(y_i, g\big(h(\boldsymbol{x}_i), s(\boldsymbol{x}_i), \boldsymbol{q}_i, \boldsymbol{w}\big)\big)$$
$$+ f(\boldsymbol{v}; \lambda), \text{ s.t. } \boldsymbol{v} \in \Psi \quad (13)$$

where $\Psi = \{\boldsymbol{v} \mid \boldsymbol{a}^T \boldsymbol{v} \leq c\}$ is a pre-defined curriculum region to initialize the weight vector $\boldsymbol{v}$. c is a constant, and $\boldsymbol{a}$ is the ranking function that indicates the difficulty levels of training samples. Usually, Ψ can be derived from a task-specific ranking function $\boldsymbol{a}$ and a constant c. In this work, $\boldsymbol{a}$ is defined by calculating $a_i = W_i^q Q_i^q$ for the ith question. W_i^q denotes the pre-defined prior weight for different question types. Q_i^q is the length of question, which is normalized by dividing the maximum question length. Note that CL is only used in the initial stage, and SPL adaptively updates $\boldsymbol{v}$ in the rest of the training stage.

IV. EXPERIMENTS

A. Datasets

To evaluate the proposed RSVQA framework, we conduct experiments on three public datasets. Two of these are released by Lobry et al. [19]: the low-resolution (LR) and high-resolution (HR) RSVQA datasets. The LR dataset is based on Sentinel-2 images at 10-m resolution. It contains 772 images of size 256×256 and 77 232 question–answer pairs. Among them, there are 23 002 (29.78%) pairs for the count question type, 22 882 (29.63%) for presence, 30 576 (39.59%) for comparison, and 772 (1.00%) for rural/urban. Overall, 77.8%, 11.1%, and 11.1% of the original tiles are divided into the training set, validation set, and test set, respectively. The HR dataset is collected from HR orthoimagery data at 15-cm resolution. It consists of 10 659 images of size 512×512 and a total number of 1 066 316 question–answer pairs. Specifically, there are 277 702 (26.04%) pairs for count, 278 335 (26.10%) for presence, 353 772 (33.18%) for comparison, and 156 507 (14.68%) for area. In general, 61.5%, 11.2%, 20.5%, and 6.8% of the original tiles are split into the training set, validation set, test set 1, and test set 2, respectively.

TABLE I
COMPARISONS ON THE LR DATASET. BOTH THE MEAN VALUE AND THE STANDARD DEVIATION ARE REPORTED

Types	Baseline	SPCL+MLL
Count	67.01% (0.59%)	**69.22%** (0.33%)
Presence	87.46% (0.06%)	**90.66%** (0.24%)
Comparison	81.50% (0.03%)	**87.49%** (0.10%)
Rural/Urban	90.00% (1.41%)	**91.67%** (1.53%)
Average Accuracy	81.49% (0.49%)	**84.76%** (0.35%)
Overall Accuracy	79.08% (0.20%)	**83.09%** (0.15%)

Another dataset is the RSIVQA dataset proposed in [43]. It is created on top of five existing datasets including UC-Merced (UCM) [51], Sydney [52], AID [53], HRRSD [54], and DOTA [55]. RSIVQA uses two types of annotation methods: manual annotation and automatic generating. In total, this dataset contains more than 110 000 VQA triplets. According to the answer type, these triplets can be divided into three types: yes/no, number, and others. Following the experimental setting described in [43], we randomly sample 80%, 10%, and 10% of all triplets as the training set, validation set, and test set, respectively.

B. Implementation Details

We use Adam optimizer with an initial learning rate of 1e-5 for model training. The batch size is set to 280 for LR and 70 for the HR dataset. For methods without multi-level feature learning, we make use of 150 epochs and 35 epochs to train the models on the LR and HR datasets, respectively. Since more epochs are needed for models with the multi-level feature learning module to converge, 300 epochs on the LR dataset and 70 epochs on the HR dataset are used for these models.

We take the method in [19] as the baseline model and conduct experiments to demonstrate the effectiveness of the proposed method. Note that the same language feature embedding network is used in both the baseline and the proposed framework. As for the visual feature learning module, we use the proposed CST and CGA modules. In addition, the traditional cross0entropy loss is replaced by the proposed SPCL loss function to enable the easy-to-hard learning strategy. To comprehensively evaluate the proposed modules, we exploit accuracy with respect to question type, average accuracy, and overall accuracy as evaluation metrics. Each model is trained three times in every experiment (except for the results in Table VI), and the mean and standard deviation are reported on both the datasets.

C. Comparisons on the LR Dataset

The proposed framework consists of two main components, namely, multi-level visual feature learning and SPCL. The numerical results on the LR dataset are shown in Table I. In this table, SPCL denotes SPCL with the soft regularizer, which is used as our final loss. SPCL + MLL represents the full-fledged framework proposed in this article. MLL stands for multi-level visual feature learning and includes both CGA and CST modules.

The LR dataset contains four types of questions: rural/urban, presence, comparison, and count. Taking into

TABLE II

COMPARISONS ON THE TEST SET 1 OF THE HR DATASET. BOTH THE MEAN VALUE AND THE STANDARD DEVIATION ARE REPORTED

Types	Baseline	SPCL	SPCL+MLL
Count	68.63% (0.11%)	68.91% (0.03%)	**69.06%** (0.13%)
Presence	90.43% (0.04%)	90.66% (0.08%)	**91.39%** (0.15%)
Comparison	88.19% (0.08%)	89.07% (0.27%)	**89.75%** (0.12%)
Area	85.24% (0.05%)	85.66% (0.26%)	**85.92%** (0.19%)
Average Accuracy	83.12% (0.03%)	83.57% (0.11%)	**83.97%** (0.06%)
Overall Accuracy	83.23% (0.02%)	83.69% (0.11%)	**84.16%** (0.05%)

TABLE III

COMPARISONS ON THE TEST SET 2 OF THE HR DATASET. BOTH THE MEAN VALUE AND THE STANDARD DEVIATION ARE REPORTED

Types	Baseline	SPCL	SPCL+MLL
Count	61.47% (0.08%)	61.63% (0.11%)	**61.95%** (0.08%)
Presence	86.26% (0.47%)	87.10% (0.12%)	**87.97%** (0.06%)
Comparison	85.94% (0.12%)	86.93% (0.23%)	**87.68%** (0.23%)
Area	76.33% (0.50%)	76.81% (0.14%)	**78.62%** (0.23%)
Average Accuracy	77.50% (0.29%)	78.12% (0.13%)	**79.06%** (0.15%)
Overall Accuracy	78.23% (0.25%)	78.25% (0.11%)	**79.29%** (0.15%)

TABLE IV

COMPARISONS ON THE RSIVQA DATASET. BOTH THE MEAN VALUE AND THE STANDARD DEVIATION ARE REPORTED

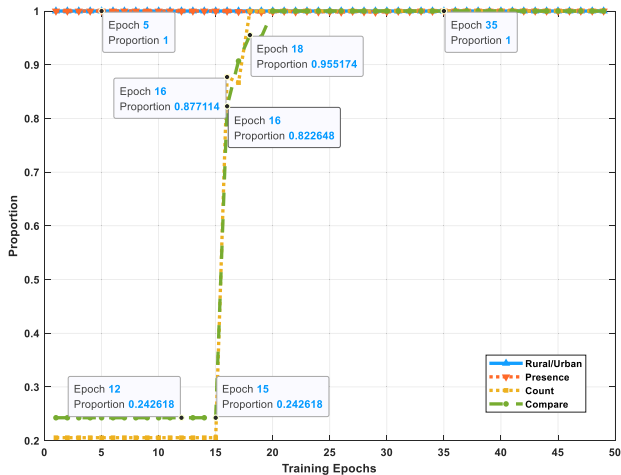| Types	MAIN [43]	SPCL+MLL
Yes/No	92.82% (0.56%)	**95.49%** (0.29%)
Number	**56.71%** (2.92%)	49.03% (0.36%)
Others	54.50% (3.06%)	**63.65%** (2.92%)
Overall Accuracy	77.39% (1.27%)	**79.70%** (0.43%)

Fig. 4. Visualization of the proportions of training samples for SPCL on the LR dataset. The proportions of different question types during the first 50 training epochs are displayed by different colors of lines.

account the difficulty level of each question type, we set the prior weights W_i^q as {rural/urban: 1.0, presence: 1.0, comparison: 3.0, count: 4.0}. From Table I, we can see that the proposed method gains better results. An improvement of 3.27% in average accuracy can be achieved. In addition, the proposed method improves the overall accuracy by 4.01%. We also find that much greater performance is obtained for all question types. Especially for comparison type, there is an improvement of about 6% in accuracy compared with the baseline method. These results indicate that multi-modal visual feature learning and SPCL training strategies can enhance the performance of RSVQA.

As shown in Fig. 4, the proportions of training samples for SPCL on the LR dataset are visualized. Note that this model does not include the MLL feature learning module. In this figure, the proportions of training samples for four question types are compared in detail. At the first 15 epochs, CL is used to initialize the weight vector v. Thus, the proportions of count and comparison question types are obviously smaller than those of rural/urban and presence. After the first 15 epochs, SPL is used to update the vector v to control the training order of different question types. The proportion of count type is the smallest at the beginning few epochs. Then, hard examples are gradually included. The results also support our assumption on the difficulty levels of different question types.

The precisions of different question types for SPCL during the training stage on the LR dataset are visualized in Fig. 5. From this figure, it can be clearly observed that easy questions can achieve higher precisions at the first 15 epochs. Afterward,
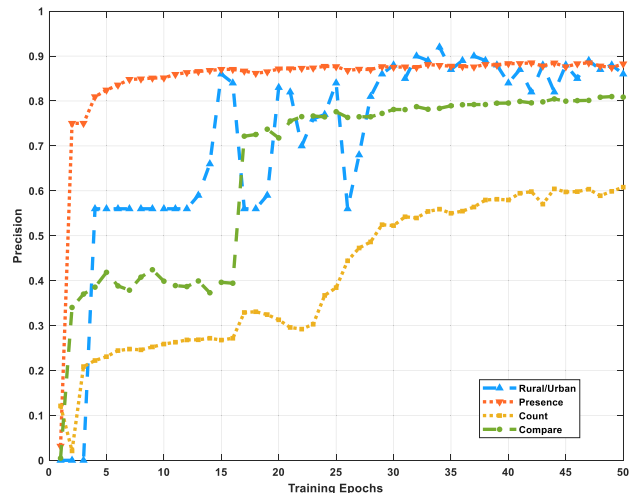

Fig. 5. Visualization of the precisions of different question types for SPCL on the LR dataset. The precisions of different question types during the first 50 training epochs are displayed by different colors of lines.

the precisions of more difficult question types are improved rapidly. Moreover, we can see that the precisions of easy question types are not affected by adding more difficult training samples.

The global attention maps and spatially transformed maps are visualized in Fig. 6. The second column of the figure shows that global attention mechanism learns to focus on important pixels of remote sensing images. The third and fourth columns indicate that spatial transformers extract visual features of local regions. Note that we apply learned transformation parameters to the original remote sensing images instead of feature maps for a clear visualization.

D. Comparisons on the HR Dataset

Since there are two test sets provided by Lobry *et al.* [19], we report the results on both of them. The experimental results on the HR dataset are displayed in Tables II and III.

Table II shows numerical results on test set 1 of the HR dataset. There are four types of questions in the HR dataset: presence, comparison, count, and
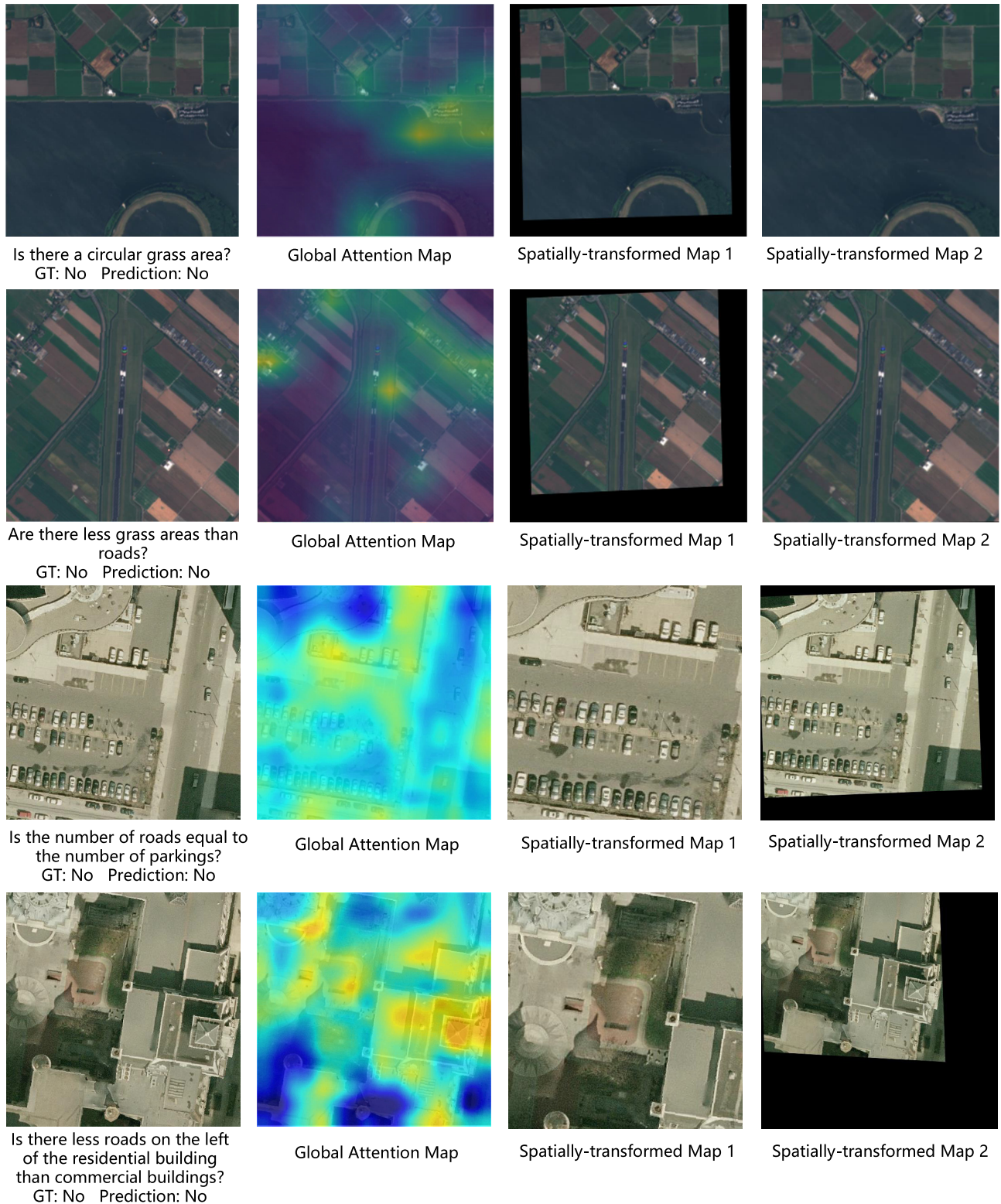
Fig. 6. Illustration of some qualitative examples on (upper two rows) LR and (lower two rows) HR datasets. The first column is the input image. The corresponding questions and predicted answers are presented below each image. Global attention maps are displayed in the second column. For the last two columns, the spatial transformation is shown (best viewed in color).

area. Their corresponding prior weights W_i^q are set as {presence: 1.0, comparison: 3.0, count: 4.0, area: 4.0}. The results in Table II reveal that SPCL can consistently enhance the performance of RSVQA for all question types. Particularly, the comparison between baseline method and SPCL indicates

that the VQA performance can be improved by simply replacing the cross-entropy loss with our SPCL loss. This demonstrates the effectiveness of the proposed training strategy.

The performance comparison between SPCL and SPCL + MLL shows that the multi-level visual feature is useful for

TABLE V
ABLATION STUDY ON THE LR DATASET. BOTH THE MEAN VALUE AND THE STANDARD DEVIATION ARE REPORTED

Types	Baseline	CGA Module	CST Module	MLL
Count	67.01% (0.59%)	67.84% (0.23%)	68.44% (0.17%)	**68.53%** (0.10%)
Presence	87.46% (0.06%)	89.49% (0.58%)	89.77% (0.22%)	**90.13%** (0.49%)
Comparison	81.50% (0.03%)	86.62% (0.33%)	**87.21%** (0.61%)	86.91% (0.50%)
Rural/Urban	90.00% (1.41%)	88.67% (1.15%)	89.00% (1.00%)	**92.00%** (1.00%)
Average Accuracy	81.49% (0.49%)	83.15% (0.38%)	83.60% (0.12%)	**84.39%** (0.26%)
Overall Accuracy	79.08% (0.20%)	81.96% (0.24%)	82.45% (0.32%)	**82.50%** (0.18%)

TABLE VI
PERFORMANCE COMPARISONS OF DIFFERENT NUMBERS OF TRAINING SAMPLES ON THE LR DATASET

Types	10% Training	40% Training	70% Training	100% Training
Count	61.22%	63.49%	67.92%	**68.61%**
Presence	80.41%	85.71%	88.85%	**90.96%**
Comparison	84.90%	85.67%	86.16%	**88.06%**
Rural/Urban	76.67%	82.50%	90.00%	**92.00%**
Average Accuracy	75.80%	79.34%	83.23%	**84.91%**
Overall Accuracy	76.50%	79.08%	81.59%	**83.23%**

this task. Comparing the baseline method with SPCL + MLL, we also find that improvements for easier question types, i.e., comparison and presence, are larger than harder ones. The experimental results reported in Table III demonstrate that the proposed method can outperform the baseline method on all question types on test set 2 of the HR dataset.

E. Comparisons on the RSIVQA Dataset

For the RSIVQA dataset, we take MAIN proposed in [43] as the baseline method. MAIN consists of two modules: a representation module and a fusion module. Image features and question representations are first learned by the representation module. Then, mutual attention and bilinear fusion are used to fuse image and question representations in an adaptive manner. The comparison between MAIN and our approach is presented in Table IV. On this dataset, according to difficulty levels of question types, prior weights W_i^q are set as {yes/no: 1.0, others: 2.0, number: 3.0}. Compared with MAIN, our proposed method can achieve much better performance on the yes/no and other types. Although the accuracy of our model on the number type is lower than that of MAIN, the proposed method can obtain better performance in general on the RSIVQA dataset. This demonstrates that the proposed training strategy and multi-level feature learning modules are effective for the RSVQA task.

F. Ablation Study and Discussion

To evaluate the proposed framework more comprehensively, the following two ablation studies are conducted to explore the effect of different sub-modules. Specifically, to show the superiority of multi-level visual features, CGA and CST modules are compared separately. Quantitative results are presented in Table V. The performance of the model with CGA on three question types is better than that of the baseline method except rural/urban type, which indicates that CGA mechanism can be used to enhance the distinguishability of

visual features. Since CST enables more flexible visual feature extraction, the model with CST can achieve better results (except rural/urban type) than both the baseline method and the model with CGA. Finally, by combining both the modules, the MLL model obtains better or competitive performance among these competing methods. This demonstrates the superiority of using multi-level visual features in the RSVQA task.

To further study the effect of different numbers of training samples on model performance, we have conducted experiments by training the proposed SPCL + MLL model with different proportions of the training data. Specifically, 10%, 40%, 70%, and 100% training samples are used for model training. The results in Table VI show that the performance of the model becomes better as the number of training samples increases gradually. In addition, we can also see that the proposed method works well with different sizes of training data.

V. CONCLUSION

In this article, two challenges for the RSVQA task are addressed. First, there are no object annotations available in the RSVQA datasets, while these annotations can provide rich semantic information for answering questions. Aiming at this challenge, a multi-level visual feature learning method is proposed to jointly learn the language-guided holistic feature and region feature. Specifically, CGA and CST modules are devised to extract flexible visual features with different spatial contexts. Second, questions in the RSVQA datasets are with clearly different difficulty levels for the same remote sensing image. Directly training a model with questions in a random order may confuse the model and limit the performance. To alleviate this problem, we propose a language-guided SPCL method with a soft weighting strategy to train networks with samples in an easy-to-hard way. Extensive experiments are conducted on three public RSVQA datasets, and the experimental results show that the proposed method can achieve state-of-the-art performance.

REFERENCES

[1] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[2] S. Talukdar *et al.*, "Land-use land-cover classification by machine learning classifiers for satellite observations—A review," *Remote Sens.*, vol. 12, no. 7, p. 1135, Apr. 2020.

[3] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," 2015, *arXiv:1508.00092*.

[4] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[5] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[6] G. Cheng, Y. Wang, S. Xu, H. Wang, S. Xiang, and C. Pan, "Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3322–3337, Jun. 2017.

[7] Z. Yuan, Q. Wang, and X. Li, "ROBUST PCANet for hyperspectral image change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 4931–4934.

[8] Q. Wang, Z. Yuan, Q. Du, and X. Li, "GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 3–13, Jan. 2018.

[9] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[10] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[11] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9962–9971.

[12] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4613–4621.

[13] T. Do, B. X. Nguyen, H. Tran, E. Tjiputra, Q. D. Tran, and T.-T. Do, "Multiple interaction learning with question-type prior knowledge for constraining answer search space in visual question answering," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 496–510.

[14] W. Zhang, J. Yu, H. Hu, H. Hu, and Z. Qin, "Multimodal feature fusion by relational reasoning and attention for visual question answering," *Inf. Fusion*, vol. 55, pp. 116–126, Mar. 2020.

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

[16] M. M. Lopez and J. Kalita, "Deep learning applied to NLP," 2017, *arXiv:1703.03091*.

[17] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," *Comput. Vis. Image Understand.*, vol. 163, pp. 21–40, Oct. 2017.

[18] Z. Yuan, L. Mou, and X. X. Zhu, "Self-paced curriculum learning for visual question answering on remote sensing data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 2999–3002.

[19] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual question answering for remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8555–8566, Dec. 2020.

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

[21] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1821–1830.

[22] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13041–13049.

[23] L. Peng, Y. Yang, X. Zhang, Y. Ji, H. Lu, and H. T. Shen, "Answer again: Improving VQA with cascaded-answering model," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 4, pp. 1644–1655, Apr. 2020.

[24] J. Yang *et al.*, "Object-centric diagnosis of visual reasoning," 2020, *arXiv:2012.11587*.

[25] Y. Zhang, J. Hare, and A. Prügel-Bennett, "Learning to count objects in natural images for visual question answering," 2018, *arXiv:1802.05766*.

[26] L. Jiang, D. Meng, Q. Zhao, S. Shan, and G. A. Hauptmann, "Self-paced curriculum learning," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2694–2700.

[27] M. Sachan and E. Xing, "Easy questions first? A case study on curriculum learning for question answering," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 453–463.

[28] Z. Xiong, Y. Yuan, and Q. Wang, "MSN: Modality separation networks for RGB-D scene recognition," *Neurocomputing*, vol. 373, pp. 81–89, Jan. 2020.

[29] H. Han, C. Li, and X. Qiu, "Multi-modal remote sensing image matching method based on deep learning technology," *J. Phys., Conf.*, vol. 2083, no. 3, Nov. 2021, Art. no. 032093.

[30] Z. Xiong, Y. Yuan, and Q. Wang, "ASK: Adaptively selecting key local features for RGB-D scene recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 2722–2733, 2021.

[31] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," 2014, *arXiv:1410.0210*.

[32] H. Kim and M. Bansal, "Improving visual question answering by referring to generated paragraph captions," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3606–3612.

[33] P. Huang, J. Huang, Y. Guo, M. Qiao, and Y. Zhu, "Multi-grained attention with object-level grounding for visual question answering," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3595–3600.

[34] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, "ABC-CNN: An attention based convolutional neural network for visual question answering," 2015, *arXiv:1511.05960*.

[35] A. Fukui, D. Huk Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," 2016, *arXiv:1606.01847*.

[36] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," 2016, *arXiv:1610.04325*.

[37] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2612–2620.

[38] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 5947–5959, Dec. 2018.

[39] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*.

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

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

[42] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, "In defense of grid features for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10267–10276.

[43] X. Zheng, B. Wang, X. Du, and X. Lu, "Mutual attention inception network for remote sensing visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[44] A. Vaswani *et al.*, "Attention is all you need," 2017, *arXiv:1706.03762*.

[45] W. Wang *et al.*, "Learning unsupervised video object segmentation through visual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3064–3074.

[46] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[47] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5209–5217.

[48] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[49] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proc. Conf. Neural Inf. Process. Syst.*, 2010, pp. 1189–1197.

[50] Q. Li, S. Huang, Y. Hong, and S.-C. Zhu, "A competence-aware curriculum for visual concepts learning via question answering," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 141–157.

[51] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th Int. Conf. Adv. Geographic Inf. Syst. (SIGSPATIAL)*, 2010, pp. 270–279.

[52] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2014.

[53] G.-S. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[54] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.

[55] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

Zhenghang Yuan (Student Member, IEEE) received the B.E. and M.E. degrees from Northwestern Polytechnical University, Xi'an, China, in 2017 and 2020, respectively. She is currently pursuing the Ph.D. degree with the Technical University of Munich, Munich, Germany.

Her research interests include remote sensing, computer vision, and deep learning.

Lichao Mou received the bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012, the master's degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2015, and the Dr.Ing. degree from the Technical University of Munich (TUM), Munich, Germany, in 2020.

In 2015, he spent six months at the Computer Vision Group, University of Freiburg, Freiburg im Breisgau, Germany. In 2019, he was a Visiting Researcher with the Cambridge Image Analysis Group (CIA), University of Cambridge, Cambridge, U.K. He is currently a Guest Professor with the Munich AI Future Laboratory "Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond (AI4EO)," TUM, and the Head of the Department "EO Data Science," Visual Learning and Reasoning Team, Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany. Since 2019, he has been the Research Scientist of DLR-IMF and the AI Consultant of the Helmholtz Artificial Intelligence Cooperation Unit (HAICU).

Dr. Mou was a recipient of the First Place in the 2016 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest and the Finalist of the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and the 2019 Joint Urban Remote Sensing Event.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition, and remote sensing.
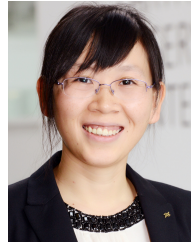
Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.Ing., and Habilitation degrees in the field of signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is currently a Professor for data science in earth observation (former: Signal Processing in Earth Observation) with TUM and the Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. Since 2019, she has been a co-coordinator with the Munich Data Science Research School, Munich (www.mu-ds.de). Since 2019, she also heads the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the Director of the International Future AI Laboratory "Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond (AI4EO)," TUM. Since October 2020, she also serves as the Co-Director of the Munich Data Science Institute (MDSI), TUM. She was the Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy; Fudan University, Shanghai, China; The University of Tokyo, Tokyo, Japan; and the University of California Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. She is currently a Visiting AI Professor with the ESA's Phi-Laboratory. Her main research interests are remote sensing and Earth observation, signal processing, machine learning, and data science, with their applications in tackling societal grand challenges, e.g., global urbanization, United Nations (UN's) sustainable development goals (SDGs), and climate change.

Dr. Zhu is a member of young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She serves in the scientific advisory board in several research organizations, among others the German Research Center for Geosciences (GFZ) and the Potsdam Institute for Climate Impact Research (PIK). She is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. She serves as an Area Editor responsible for special issues for *IEEE Signal Processing Magazine*.