

# Visual Question Generation From Remote Sensing Images

Laila Bashmal, *Graduate Student Member, IEEE*, Yakoub Bazi<sup>✉</sup>, *Senior Member, IEEE*, Farid Melgani<sup>✉</sup>, *Fellow, IEEE*, Riccardo Ricci<sup>✉</sup>, *Graduate Student Member, IEEE*, Mohamad M. Al Rahhal<sup>✉</sup>, *Senior Member, IEEE*, and Mansour Zuair<sup>✉</sup>, *Member, IEEE*

**Abstract**—Visual question generation (VQG) is a fundamental task in vision-language understanding that aims to generate relevant questions about the given input image. In this article, we propose a paragraph-based VQG approach for generating intelligent questions in natural language about remote sensing (RS) images. Specifically, our proposed framework consists of two transformer-based vision and language models. First, we employ a swin-transformer encoder to generate a multiscale representative visual feature from the image. Then, this feature is used as a prefix to guide a generative pretrained transformer-2 (GPT-2) decoder in generating multiple questions in the form of a paragraph to cover the abundant visual information contained in the RS scene. To train the model, the language decoder is fine-tuned on RS dataset to generate a set of relevant questions from the RS image. We evaluate our model on two visual question-answering (VQA) datasets in RS. In addition, we construct a new dataset termed TextRS-VQA for better evaluation for our VQG model. This dataset consists of questions completely annotated by humans which addresses the high redundancy of the questions in prior VQA datasets. Extensive experiments using several accuracy and diversity metrics demonstrate the effectiveness of our proposed VQG model in generating meaningful, valid, and diverse questions from RS images.

**Index Terms**—Language transformers, remote sensing (RS), vision, and visual question generation (VQG).

## I. INTRODUCTION

**D**URING the last decade, the capabilities of remote sensing (RS) technologies have drastically improved, leading to the availability of valuable data for earth monitoring applications such as urban planning and change detection. However, as the advances in the RS acquisition platforms have grown, so has

the need to build appropriate techniques for extracting useful knowledge from such data. Early techniques for RS image interpretation are mainly proposed to solve the classical tasks of scene classification and object detection which aim to recognize the image as a whole or the different land-cover classes in the images. However, these techniques are inadequate to express the rich and complex information in the RS scene as they ignore the attributes and the high-level relationship between objects present in the scene.

The long-standing objective of automatic RS understanding is to provide a thorough, human-like interpretation of the RS images that common users who may lack technical knowledge in RS can easily understand [1]. Since language is the most convenient mean for human to communicate with the world, it becomes essential to incorporate natural language processing (NLP) algorithms into the automatic understanding of the remotely sensed data. Vision-language integration is currently gaining a lot of attention from researchers and it is actively practiced in a variety of tasks in the RS such as describing the content of RS image via image captioning [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], retrieving RS image using textual query [17], and more recently, visual question answering (VQA) [18], [19], [20], which learns the machine to answer a question formulated in natural language about the image.

Visual question generation (VQG) represents the natural extension to the aforementioned tasks. This task aims to generate questions in natural language about the content of the image. VQG is an innovative but challenging cross-disciplinary problem, which requires both computer vision techniques to analyze the content of the image and NLP methods to produce the question. Enabling the machine to ask relevant and semantically coherent questions has broad applications. It is an integral part for any interactive system. Either to initiate a conversation with the end-user, or to inquiry about specific information to accomplish a certain task. Additionally, exploiting the relationship between VQA and VQG is a step toward building a conversational VQA dialog system for describing the image. Automatic questions generation can also help in automating the annotation of the VQA datasets. For these advantages, great progress has been made in generating questions from natural scenes [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], and from images in specific domain such as medical images [35]. However, similar problem has not yet been explored in the RS.

Manuscript received 25 October 2022; revised 25 December 2022 and 23 February 2023; accepted 14 March 2023. Date of publication 24 March 2023; date of current version 7 April 2023. This research was supported by the researchers supporting project number (RSPD-2023R607), King Saud University, Riyadh, Saudi Arabia. (Corresponding author: Yakoub Bazi.)

Laila Bashmal, Yakoub Bazi, and Mansour Zuair are with the Computer Engineering, King Saud University, Riyadh 145111, Saudi Arabia (e-mail: 439204359@student.ksu.edu.sa; ybazi@ksu.edu.sa; zuair@ksu.edu.sa).

Farid Melgani is with the Department Computer Science and Information Engineering, University of Trento, 38123 Trento, Italy (e-mail: melgani@disi.unitn.it).

Riccardo Ricci is with the Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: riccardo.ricci-1@unitn.it).

Mohamad M. Al Rahhal is with the College of Applied Computer Sciences, King Saud University, Riyadh 11451, Saudi Arabia (e-mail: mmalrahhal@ksu.edu.sa).

The TextRS-VQA dataset is available at: <https://github.com/yakoubbazi/TextRS>.

Digital Object Identifier 10.1109/JSTARS.2023.3261361

VQG is a novel problem in RS. Therefore, some unique challenges need to be considered in this field. Among the main challenges is the lack of a high-quality dataset for question generation. One viable solution is to rely on the publicly available datasets proposed for VQA, which consists of image-questions-answers triplets such as the RSVQA-LR [18] and the RSIVQA-DOTA [20]. The RSVQA-LR [18] dataset is annotated with around 100 questions per image of five types (count, presence, area, comparison, and rural/urban), and the RSIVQA-DOTA [20] provides 3-24 question-answer per image of two basic types, count and presence questions. Although these datasets have contributed in advancing VQA methodologies in RS, they all share the same limitations. These datasets are built to be answer-centric. In other words, the answers in these datasets are diverse as they are derived from different scene classification and object detection datasets. However, the questions are automatically generated from predefined templates, which results in high redundancy in the questions. For example, 89% of the questions in RSVQA-LR are redundant, and in RSIVQA-DOTA, out of 16 430 questions, only 32 of these questions are unique. This can have a negative impact on the question generation results.

Besides the datasets challenge, RS images are often characterized by their abundant and complex visual contents. Therefore, generating a single question is often insufficient to capture all the important information presented in the image. Different from VQA task where the answer to the question is typically deterministic, in VQG multiple valid questions could be formed from objects, attributes, and/or relationships. Ideally, a VQG system should be able to mimic the ability of humans in analyzing a complex image and generating multiple questions about different concepts within the image. However, generating various questions while maintaining a decent balance between the correctness and the diversity of the generated questions is a key challenge.

In light of the above-mentioned observations, we introduce VQG model for RS images to provide intelligent questions in natural language about the image. Our model is a paragraph-based that emulates the human ability in generating a series of questions to cover all the information contained in the RS scene. Specifically, we utilized a vision transformer model known as a swin transformer to learn representative visual features from the image. The hierarchical architecture of this model along with the self-attention mechanism enables it to effectively extract global multiscale feature representations from the RS image. This feature representation is fed as a token to the language model to guide the generation of the questions. For the language model, we leverage the power of the generative pretrained transformer-2 (GPT-2) model, which is an auto-regressive model built upon the transformer. This language model is fine-tuned on the RS dataset to generate a set of the questions from the image in the form of a paragraph. In addition, we present the TextRS-VQA dataset, a new manually annotated dataset that has been constructed for VQA and VQG task in RS. Unlike prior datasets that have high redundancy due to the automatic generation of the questions, this dataset is fully annotated by human annotators.

Overall, the main contributions of this work are summarized as follows.

- To the best of our knowledge, this work is the first to introduce the VQG task in the RS domain.
- We propose a question generation model consisting of two transformer-based vision and language models. The vision model generates a visual representation, which is then used by the language model to generate the questions.
- We introduce a new VQA dataset for RS data termed TextRS-VQG dataset to advance the task of VQG in RS. This dataset is manually annotated and distinguished from other datasets by having diverse questions of various types.
- We evaluate our method on two VQA datasets along with the proposed TextRS-VQA dataset. The results indicate that our model performs well at generating relevant, diversified, and meaningful questions.

The rest of this article is organized as follows. In Section II, we review the related works. Section III introduces the proposed VQG model. Section IV presents our TextRS-VQA dataset. Section V presents the experiment setup and the results of the experiments. Finally, Section VI concludes this article.

## II. RELATED WORK

In the last few years, significant progress in a variety of vision-and-language tasks has been made both in computer vision and RS communities. In this section, we review research related to VQG task. Specifically, VQA in RS, VQG in computer vision and RS image captioning.

### A. VQA

VQG is closely related to the literature of VQA, which is the task with the objective of giving an answer given an image and a question posed in natural language about the image. VQA has recently been introduced in the field of RS. Existing models are commonly composed of a visual feature encoding model for the image and a language encoding model to encode the question. The output features of the two models are then combined and fed into an answer prediction model. The task of VQA in RS has first been proposed by Lobry et al. [18], in which a VQA model is composed of a convolutional neural network (CNN) for feature extraction and a recurrent neural network (RNN) for question encoding. Then, both the visual information and the question representation are fused to obtain a single feature vector. Finally, a classifier network is used for predicting the answer. In a subsequent work, Zheng et al. [20] proposed a model that enhanced the feature representation in [18], by including a mutual attention mechanism in the fusion step. The features of the image and the question are fused through a bilinear layer to allow the model to consider the semantic correspondence between the image and the question. In a different but related work, Yuan et al. [19] proposed a VQA model for change detection in multitemporal RS images in which the model can provide answers in natural language about the difference between two RS images. The model consists of two CNNs for encoding multitemporal images, a multitemporal fusion block to fuse the features of the two images, a multimodal fusion block for fusing the vision features

with the textual question, and finally, an answer prediction block. In [36], different fusion strategies for the image and the question are evaluated. The results indicate that more complex fusion strategies yield higher accuracies. More recently, Siebert et al. [37], proposed a VQA model that uses a multimodal fusing module based on VisualBERT to integrate the image and the language modalities.

### B. VQG

VQG has been an active research area in the computer vision community. The goal of the VQG is to learn the machine to ask questions formulated in natural language about the content of the image. This task was introduced in its current form in [21], in which the image and a generated caption from the image are used to generate the questions. Since then, a large number of studies has tackled this problem from different perspectives. For example, Mostafazadeh et al. [22], proposed a VQG model with the goal of generating more engaging and natural questions rather than literal questions about the image. In [23], a model based on BERT model is proposed. The model utilized both object visual features and image captions for question generation.

Since multiple valid questions can exist for a single image, follow-up methods in the literature proposed to generate diverse questions from the given image, either by using generative models such as variational auto-encoders (VAEs) [24] or search methods such as diverse beam search [25].

On the contrary, some works on VQG attempt to control the generation of the question by conditioning the model on an auxiliary information such as the ground-truth answer [26], [27], [28], or the type of the expected question [30], [31], [32], [33]. Conditioning the generation of the question on the answer is known as the inverse VQA, in which the model is given an image and an answer pair to generate the appropriate question. Liu et al. [26] proposed a model based on a CNN for encoding the image, a long short-term memory (LSTM) for encoding the answer, and another LSTM with an attention module to generate the question. This model has been improved in [27] by integrating a VAE to generate diverse questions from generic answers such as “yes/no” and numbers. Alwattar and Guo [28] proposed a model composed of a gated recurrent unit (GRU) to encode the answer and a faster R-CNN to generate visual features at object level. The model employed a multilevel attention module to enhance the visual features with their corresponding text before generating the question with a GRU decoder. Li et al. [38] leveraged the complementary relationship between the VQA and the VQG and proposed a model that can accomplish the two tasks together. When the model is given the question, the model generates the answer and vice versa.

Alternatively, other works avoid using the answer as prior information, as this contradicts the flow of the natural conversation. As an example, Krishna et al. [33] proposed to sample the latent variable model using the question type distribution instead of the answer to help generate the corresponding questions. Furthermore, Vedd et al. [30] proposed a model which conditions the question generation on the question type, or the objects present in the scene. This conditional information could

be explicitly selected by the user or automatically induced by the model. In [31], VAE is used to generate multiple questions per image of the specified question type. A similar model is proposed in [32] but with improved consistency between the image, the question, and the question type by training the model on a consistent cyclic loss.

Another line of work in VQG aims to generate not only the question but also the expected answer from the image. In [34], a model composed of a CNN-LSTM is proposed to generate both the question and answer from the image.

### C. Image Captioning

VQG task is related to the task of image captioning. Both tasks take an image as an input and generate an output in natural language. However, image captioning aims to generate descriptive sentence about the content of an image instead of a question. Earlier approaches for generating captions in RS are either template-based [5] or retrieval-based approaches [6]. These approaches can generate captions that are grammatically correct, but they are often simple and redundant. Thus, works based on the encoder-decoder paradigm have become the mainstream in RS image captioning due to the quality and the diversity of the captions they can generate. In this paradigm, image features are encoded via a vision model and a sentence is generated by decoding those features using a text model. The work of Qu et al. [7] is among the first works that have followed this paradigm. In which, a pretrained CNN extracts the image features, and an LSTM is used as a decoder to generate the caption. In [2], multiscale features from different layers of CNN are fused to deal with the scale variation of objects in RS images. Fu et al. [3] added an external memory to the LSTM to generate a more comprehensive caption. Sumbul et al. [4] proposed to use a summarization module to merge redundant ground truth captions into one, which is then used to train the captioning model. Hoxha and Melgani [8] proposed a method for training a captioning model with limited training samples. The method used a decoder of multiple support vector machines to alleviate overfitting. In [9], a CNN is used to generate multiple labels. Then, valuable labels are selected and used with the ground truth captions to train the model. Li et al. [10] proposed to optimize the model on a truncation cross-entropy loss instead of the regular cross-entropy loss to make the model less sensitive to overfitting. In [11], several models based on an encoder-decoder are explored, and the attention mechanism is used to enhance feature representation. Zhang et al. [12] introduced an attribute attention module so that the model can pay more attention to the important regions of the image while generating the caption. In [13], a model with three levels of attention is used. The attention to different areas of the image, the attention to different words, and the attention to vision and text. Yuan et al. [14] proposed a model with a two-branch encoder. The first uses a CNN with multilevel attention and the second uses graph CNN to utilize visual relationships within the RS image. In [15], a model used a CNN to predict the label, which is then utilized to guide the calculation of the attention mask. Finally, Zhang et al. [16] proposed to improve the attention mechanism by learning

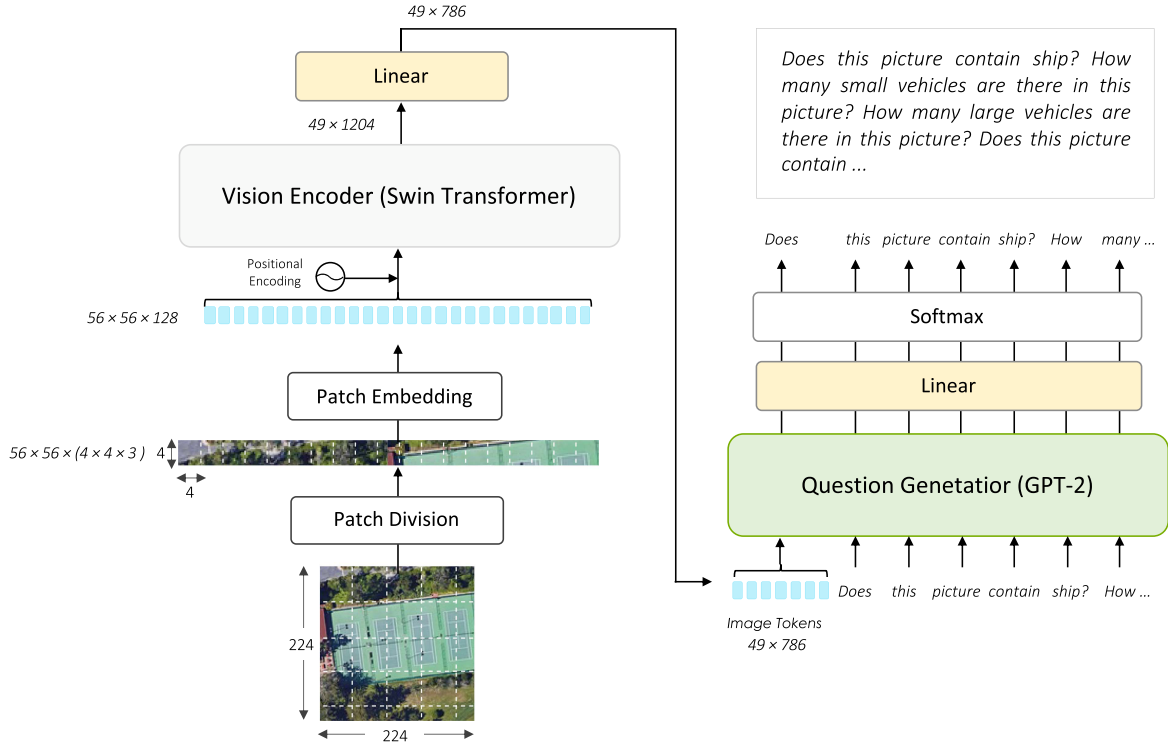


Fig. 1. Overview of our VQG model. We first use the vision encoder to obtain the image tokens  $f_i$ . These tokens are then passed to the question generator to get the questions for the image in a form of a paragraph.

the model on a visual aligning loss after filtering out nonvisual words.

### III. PROPOSED MODEL

Let  $\{\mathbf{X}_i, \mathbf{Q}_i\}_{i=1}^I$  be a dataset of  $I$  image and question pairs. Each RS image  $\mathbf{X}_i$  is annotated with one or more questions  $\mathbf{Q}_i = \{q_{i,j}\}_{j=1}^J$  where  $J$  is the number of questions associated with that image. We derive the question paragraph  $R_i$  by sequentially concatenating the questions in  $\mathbf{Q}_i$

$$R_i = q_{i,1}; q_{i,2}; \dots; q_{i,j}. \quad (1)$$

The questions paragraph  $R_i$  can be formulated by the sequence of words that composes all the questions in the paragraph  $R_i = \{w_{i,k}\}_{k=1}^K$ , where  $w_{i,k}$  is the  $k$ th word in the paragraph and  $K$  is the total number of words in that paragraph.

The goal of our paragraph-based VQG is to use the information present in the RS image to generate meaningful, valid, and diverse questions in natural language.

The overall architecture of our proposed VQG model is illustrated in Fig. 1 which consists of two parts: first, vision encoder which generates representative visual features from the RS image, and second, question generator for generating a paragraph of questions given the visual tokens. In this work, we employ a pretrained vision transformer model as an image backbone, and a GPT-2, which is a pretrained paragraph-based generative model as the questions generator.

#### A. Vision Encoder

The first step in our VQG model is to represent the RS image with representative visual features. Since generating questions requires fine-grained image understanding, we make use of a variant of the vision transformer model known as swin transformer [39]. This model is built upon vision transformers, which has shown a remarkable performance on different image tasks [40]. The main motivation for this choice is the hierarchical architecture of the swin transformer model which allows it to recognize visual elements of different scales, and the self-attention mechanism that enables it to capture global dependencies within the image.

Given an RS image  $X_i$  of dimension  $224 \times 224 \times 3$ . The image is first divided into tiny nonoverlapping patches of size of  $4 \times 4 \times 3$  to form a sequence of patches of dimension  $56 \times 56 \times 48$ . Each patch in this sequence is treated as a token. The sequence is then projected by an embedding layer into the vision encoder dimension to form a feature of size  $56 \times 56 \times 128$ . The positional information is added to the sequence and then passed to the vision encoder.

As shown in Fig. 2(a), the vision encoder consists of multiple swin transformer blocks and patch merging layers that work together to generate the hierarchical visual feature. The swin transformer block is responsible for learning the feature representation, while the patch merging layer is responsible for reducing the number of tokens between the swin transformer blocks by merging adjacent patches. The input to each block is first partitioned into windows, each contains  $7 \times 7$  patches. Then, two configurations of the multihead self-attention (MSA) are

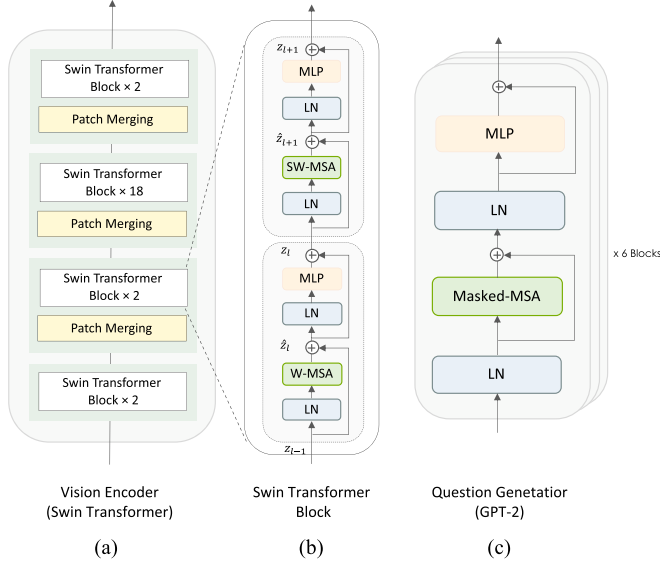


Fig. 2. Internal architecture of the (a) vision encoder, (b) swin transformer block, and (c) GPT-2 block.

employed. The window-based (W-MSA) which computes self-attention locally within each window, and the shifted window-based (SW-MSA) which shifts the features before partitioning and computing the attention. The local computation within the window employed by the first configuration reduces the complexity of the self-attention mechanism, while the shifted window-based self-attention allows to model long-range cross-window dependencies. Both configurations of the MSA use the self-attention mechanism. First, the input feature is transformed into three independent features, i.e., the query  $Q$ , key  $K$ , and value  $V$  using three linear layers. Then, these features pass through multiplication, scaling, and a softmax layer to generate the attention vector which is computed as follows:

$$Attention = softmax \left( \frac{QK^T}{\sqrt{d}} + B \right) V \quad (2)$$

where  $Q, K, V \in R^{M^2 \times d}$  are the query, the key, and the value matrices, respectively,  $d$  is the dimension of the key,  $B$  is a learnable relative positional encoding and  $M$  is the window size. In MSA, the attention function is performed by multiple heads in parallel and the results are concatenated to be fed as an input to the next layer.

As shown in Fig. 2(b), in addition to the two MSA layers, the swin transformer block has four normalization layers (LN), two multilayer perceptron (MLP) networks, each has two linear layers with a GeLU activation function. The different modules of swin transformer block are connected via skip connections. The layers inside each swin transform block can be expressed as follows:

$$\hat{z}_l = WMSA(LN(z_{l-1})) + z_{l-1} \quad (3)$$

$$z_l = MLP(LN(\hat{z}_l)) + \hat{z}_l \quad (4)$$

$$\hat{z}_{l+1} = SWMSA(LN(z_l)) + z_l \quad (5)$$

$$z_{l+1} = MLP(LN(\hat{z}_{l+1})) + \hat{z}_{l+1}. \quad (6)$$

After passing through several patch merging layers and swin transformer blocks, we utilize the output of the last block before the average pooling layer as the image feature representation, which has the dimension of  $49 \times 1204$ . Before feeding the features into the language model, we project it by a fully connected layer into the dimension  $49 \times 786$

$$f_i = FC_1(z_L) \quad (7)$$

where  $f_i$  represents the hierarchical visual feature associated with image  $X_i$  and  $z_L$  is the output of the last transformer's block. The feature vector  $f_i$  is then fed into the language model as a prefix for generating the questions.

## B. Questions Generator

The main goal of the questions generator is to produce multiple questions for the given RS image. Given the success of the GPT models in general text generation tasks, we propose to employ GPT-2 [41] for generating questions. GPT-2 is a paragraph-based generative model created by OpenAI in 2019 for generating long paragraphs.

There are two motivations for our choice. First, the GPT-2 architecture is based on the language transformer decoder [42], which has proven to be more effective than sequential models such as RNN and LSTM in many language modeling tasks. Second, GPT-2 was pretrained on a large-scale corpus of 8 million web pages and has over of 1.5 billion parameters. This is useful when training our model on RS datasets which are relatively small and costly to annotate. We adapt the GPT-2 model to generate questions from RS imagery by fine-tuning it on RS datasets. This allows transferring the linguistic knowledge learned in the GPT-2 to the RS questions generation task.

Specifically, GPT-2 generates a paragraph one word at a time in an auto-regressive fashion. It predicts the next word conditioned on a sequence of the prior tokens. Then, when the word is predicted, that word is added to the input sequence and the new sequence becomes the input to the model in its next step. In our case, the questions are generated conditioned on the features of the given RS image. We feed the language model with the visual vector  $f_i$  as a prefix. Then, the model learns to generate a paragraph of questions word after word starting from the visual tokens.

The architecture of the adopted question generation model is the distilled version of the GPT-2 termed as DistilGPT2, which has six identical transformer blocks. The DistilGPT2 can take up to 1024 token length and has 768-hidden feature representation, 12 attention heads, and 82M parameters (compared to 124M parameters for GPT-2). The internal architecture of a single GPT-2 block is shown in Fig. 2(c). The GPT-2 uses the masked-MSA layer that applies self-attention on the previous tokens by masking the future positions. The block also contains an MLP, and two normalization layers connected via skip connections just like the original transformer.

The question generator is followed by a linear layer that projects the output produced by the stack of the question generator decoders into a logits vector. This logits vector is turned into probabilities by softmax layer, where the word with the

highest probability is chosen to be the next word in the question paragraph.

### C. Training and Loss Function

The goal of training is to fine-tune the parameters of the questions generator using the labeled RS training data, which consists of pairs  $(X_i, R_i)$  of an image  $X_i$  and the corresponding paragraph of questions  $R_i$ . During training, the model receives as input an RS image  $X_i$ . The vision encoder generates the visual feature representation from the image. The objective of the question generator is the auto-regressive language modeling objective function conditioned on the feature representation of the image  $X_i$  and the past tokens of the question paragraph. Formally, the objective function is defined as follows:

$$\mathcal{L} = - \sum_{k=1}^K \log P(w_{i,k} | w_{i,0} \dots w_{i,k-1}, f_i) \quad (8)$$

where  $f_i$  is the feature representation of the image  $X_i$ ,  $w_{i,0}, \dots, w_{i,k-1}, w_{i,k}$  are the words composing the questions of the paragraph  $R_i$  associated with the image  $X_i$ , and  $K$  is the number of words in the paragraph.

At the inference phase, the test image is given as an input to the VQG model. The model generates questions one word at each time step by sampling the word that has the highest probability. The predicted words constitute the paragraph of questions corresponding to the test image.

## IV. DATASETS

VQG has made significant progress in computer vision. A key reason for this progress is the availability of large datasets with high-quality questions of various types [22], [43]. However, in the RS domain, the development is still limited, and the existing datasets such as RSVQA [18] and RSIVQA [20] which were proposed to solve the VQA problem are focused more on the answers rather than the questions. The questions in these datasets are generic, repetitive, and lacking in diversity and flexibility as they are automatically generated from predefined templates. Thus, it is essential to establish a VQG benchmark to fill the gap and advance the task of question generation in the RS domain. Therefore, we construct the TextRS-VQA dataset, which is the first complete manually annotated dataset for VQA and VQG tasks in the RS domain. This dataset addresses the problem of high redundancy of the template-based questions in the prior datasets. In the following, more information about the proposed dataset and the other VQA datasets in RS is presented.

### A. TextRS-VQA Dataset

This dataset is based on the images of the TextRS dataset which was introduced for RS image retrieval and captioning [17]. It was built by collecting images from four well-known scene classification datasets with different image sizes and spatial resolutions. Namely, AID [1], PatternNet [45], UC-Merced [46], and NWPU [47]. From each dataset, 16 random images are extracted from every class. This creates a new dataset with a total number of 2144 of images, each image is annotated manually

with two to five questions to ensure diversity, and the given questions are more tailored to RS use cases. Our dataset has 6245 questions of four types which are: object counting, presence/absence, class type, and the “other” type which includes a wide range of questions. We select 80% of the images and their associated questions for training while 20% are left for test. Since this dataset is manually annotated it has the most diversity ratio with more than 57% unique questions.

### B. VQA Datasets in RS

In this section, we describe the other VQA datasets used for evaluating our VQG model, which are: RSVQA-LR and RSIVQA-DOTA datasets. These datasets are characterized by low diversity in the questions and small questions vocabulary size.

1) *RSVQA-LR* [18]: This dataset was originally created for VQA in RS. The annotation for this dataset was automatically generated from predefined templates. The information used to construct the question-and-answer pairs is derived from OpenStreetMap. In this work, we used the low-resolution subset which includes four types of questions (object counting, comparison, presence/absence, rural/urban). The images in this dataset consist of 9 different tiles covering an area of 6.55 km<sup>2</sup> over The Netherlands. The images were acquired using Sentinel-2 satellite with a spatial resolution of 10 m. Each image in this dataset has the size of 256 × 256 with RGB spectral channels. The dataset contains 772 images, the default splitting of the dataset is 572, 100, and 100 images for training, validation, and testing, respectively. Each image is annotated with 100-101 questions, which makes it the highest dataset in terms of the number of questions per image. However, since this dataset was acquired over one location and the questions are automatically generated, many of these questions are redundant and only 11% of these questions are unique. Thus, we selected only the first 50 questions for each image.

2) *RSIVQA-DOTA* [20]: This VQA dataset is derived from existing RS scene classification and object detection datasets. Most of the questions are automatically generated using scene and object level annotation, with a small part of the dataset being annotated by humans. In this work, we specifically used the part derived from the DOTA [48] object detection dataset. Each image in the dataset is labeled with 3-24 questions about the presence of the object and the number of objects in the scene. The total number of questions given to the image is 16 430, but only 32 of these questions are unique. The dataset is split into two sets the training set, which contains 1298 images and the test set, which contains 260 images.

Fig. 3 shows some sample images and their associated questions from our datasets and the previously proposed VQA datasets in RS. The figure illustrates that the questions in our dataset are more natural and are highly relevant to the image’s content. It is worth mentioning that the answers provided in these datasets are unnecessary for the targeted VQG problem. Furthermore, Table I summarizes some statistics from our TextRS-VQA dataset and the other two VQA datasets in RS. The table shows that although our dataset has the lowest questions per image

TABLE I  
COMPARISON BETWEEN VQA DATASETS IN RS

Dataset	#Images	#Questions	#Unique questions	Diversity	#Questions per image	Avg. question length	Question vocabulary size
TextRS-VQA	2143	6245	3608	57.77%	2-5	6.62	609
RSVQA-LR	772	77232	8206	10.62%	100-101	8	131
RSIVQA-DOTA	1868	16430	32	0.002%	3-24	6.67	47

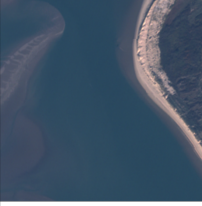
RSVQA-LR	RSIVQA	TextRS
		
<ul style="list-style-type: none"> <li>• Is it a rural or an urban area?</li> <li>• Is the number of commercial buildings equal to the number of commercial buildings?</li> <li>• Is there a farmland?</li> <li>• ...</li> </ul>	<ul style="list-style-type: none"> <li>• How many small vehicles are there in this picture?</li> <li>• Does this picture contain small vehicle?</li> <li>• Does this picture contain plane?</li> <li>• ...</li> </ul>	<ul style="list-style-type: none"> <li>• Is there aircrafts next to the loading area?</li> <li>• What is the green area next to the terminal?</li> <li>• How many aircrafts in this terminal?</li> </ul>
		
<ul style="list-style-type: none"> <li>• Is a commercial building present?</li> <li>• Is there a building?</li> <li>• How many buildings are there?</li> <li>• What is the number of roads?</li> <li>• Is a water area present?</li> <li>• Are there more forests than buildings?</li> <li>• Is there a small island?</li> <li>• ...</li> </ul>	<ul style="list-style-type: none"> <li>• How many baseball diamonds are there in this picture?</li> <li>• How many soccer ball fields are there in this picture?</li> <li>• Does this picture contain baseball diamond?</li> <li>• Does this picture contain plane?</li> <li>• Does this picture contain harbor?</li> <li>• ...</li> </ul>	<ul style="list-style-type: none"> <li>• How many basketball courts are in the image?</li> <li>• Are there buildings in the image?</li> <li>• Is there a green passage near the basketball courts?</li> </ul>

Fig. 3. Example of images and the corresponding questions from the datasets: TextRS-VQA, RSVQA-LR, and RSIVQA-DOTA.

rate, its diversity, as measured by the proportion of unique questions to all other questions in the dataset, is higher. In addition, compared to the other datasets, our dataset has the shortest average question length, but a richer vocabulary.

Fig. 4 visualizes the distribution of the question types for the three datasets. As can be seen from the figure, the RSVQA-LR has four question types and the RSIVQA-DOTA has only two. On the other hand, in the TextRS-VQA dataset, roughly half of the questions are from the “presence,” “class type,” and “count” types, and the other half has a mix of questions. To further illustrate the diversity of our dataset, Fig. 5 provides an in-depth analysis of the questions. The frequency distribution of the first three words of the questions in each dataset is displayed using a sunburst visualization. As can be seen, when compared to the other VQA datasets, our TextRS-VQA dataset revealed a higher diversity.

## V. EXPERIMENTAL RESULTS

In this section, we first provide the experimental details of our VQG model and explain the evaluation metrics utilized in this work. Lastly, we present the model’s results.

### A. Experimental Details

We used the (Swin-Base) version of the swin transformer as a vision backbone. The model was pretrained on ImageNet-22k and fine-tuned on ImageNet-1k dataset at resolution  $224 \times 224$ . The model uses a patch size of  $4 \times 4$  pixels and a window size of  $7 \times 7$  patches. The model has an embedded dimension  $C$  of size 128. The base model consists of four layers, the first, the second and the last layers each has two swin transformer blocks, while the third layer has 18 blocks. The patch merging is applied after the first three blocks. For the question generation, we used a distilled version of the GPT-2 termed as DistilGPT2. This model can take up to 1024 token length and has 6 layers, 768-hidden feature representation, 12 attention heads, and 82 million parameters.

When fine-tuning the language model, we used Adam [49] for optimization with mini-batch size of 16 and 128 maximal epochs. The initial learning rate is set to  $2e-5$ , which decays at a rate of 0.05. For all datasets, we used a sequence length of 400 for the GPT2 model. Sequence shorter than this limit is padded with zero.

The number of questions in the paragraph is selected based on the number of questions per image in each dataset. The top 20 generated questions are considered for the RSVQA-LR and the RSIVQA-DOTA datasets. However, since the TextRS-VQA dataset has fewer number of questions per image, we considered only the top 10 questions.

### B. Metrics

Different evaluation metrics are utilized to assess the quality of the questions generated by our VQG model. The first group of metrics are the similarity metrics, which evaluate how well the generated questions match with the ground-truth questions in the dataset.

The first similarity metric is the bilingual evaluation understudy (BLEU) [50] score, which measures the co-occurrences of the matching words (n-gram) between the generated question and the ground-truth question, where n-gram is a set of one or more ordered words and it is chosen to be between 1–4. The second is the metric for evaluation of translation with explicit ordering (METEOR) score [51], which is computed by building an alignment between the words in the predicted

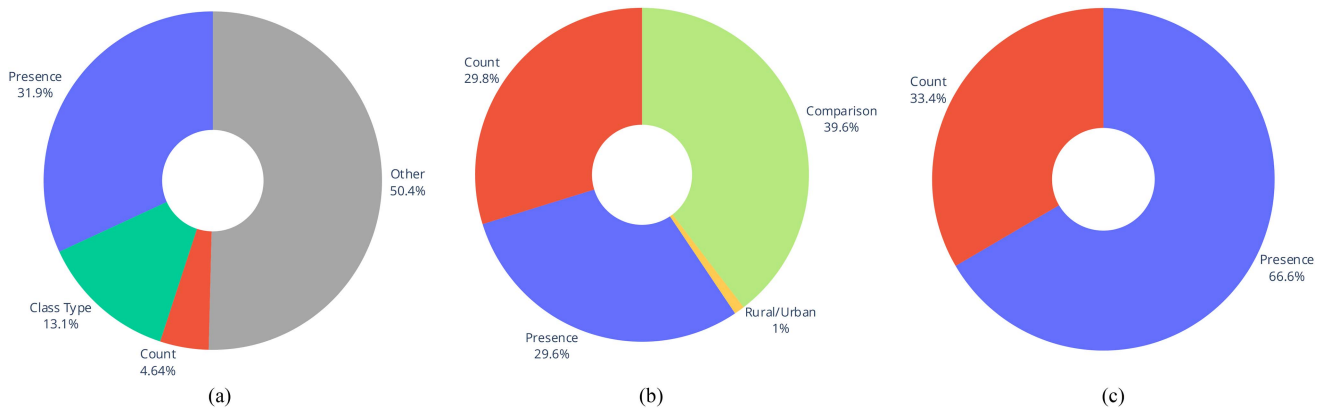


Fig. 4. Distribution of the question type for the datasets. (a) TextRS-VQA. (b) RSVQA-LR. (c) RSIVQA-DOTA.

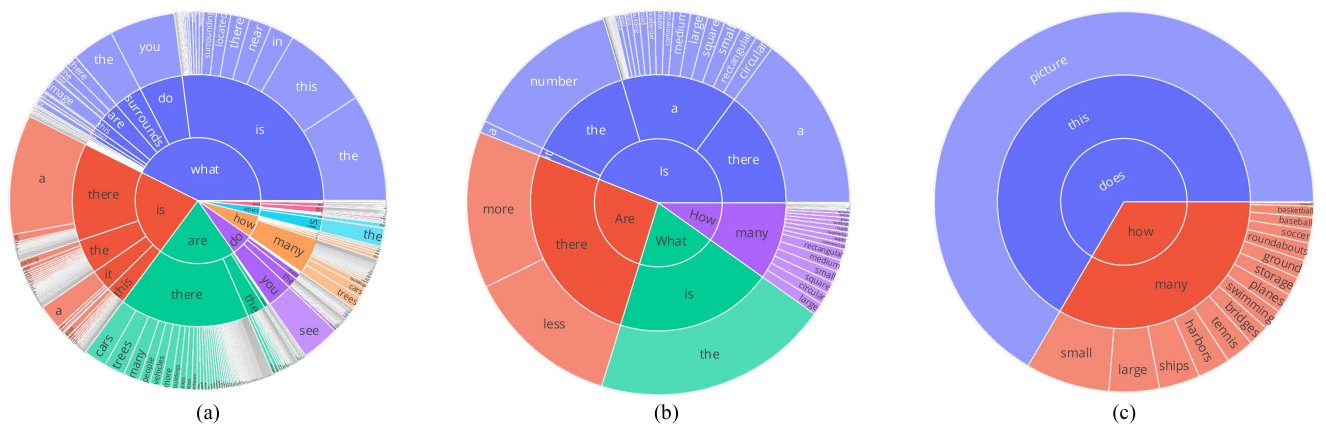


Fig. 5. Sunburst visualization for the datasets. (a) TextRS-VQA. (b) RSVQA-LR. (c) RSIVQA-DOTA.

question and the ground-truth questions. The consensus-based image description evaluation (CIDEr) score [52] measures the similarities by adding a term frequency–inverse document frequency (TF-IDF) for every n-gram. This metric reduces the weight of nonkeywords and words with high frequency. Additionally, we computed the recall-oriented understudy for gisting evaluation (ROUGE). This metric quantifies the similarity of the longest common subsequence between generated and ground-truth questions. The range of all the above-mentioned metrics except the CIDEr is between zero and 100, the closer to 100 the metrics are, the more likely the generated question is similar to the reference question in the dataset.

Additionally, another type of metrics is employed to assess the diversity of the generated questions which includes the strength and inventiveness as described in [24]. The strength is defined as the proportion of unique questions generated per image, while the inventiveness is defined as the ratio of the original generated questions never seen in the training set.

### C. Results of Similarity Metrics

First, we quantitatively verify the performance of our question generation model in terms of the similarity metrics discussed previously. Fig. 6 shows the results of our proposed VQG on the

three datasets. The x-axis represents the order of the question in the paragraph, while the y-axis represents the values of the similarity metrics. Table II shows the scores of the first and the last generated questions in the paragraph.

In general, the results of the TextRS-VQA are lower compared to the other two datasets, especially in the BLEU scores, ROUGE, and METEOR. This is because the questions in our dataset are manually labeled and more challenging for the VQG model to generate. The results of the TextRS-VQA show a decrease of around 5%–11% in almost all scores between the first and the last generated question in the paragraph. The CIDEr score shows a rapid decrease from 85.64% to 48.85%.

The results of the RSVQA-LR show that the first question in the paragraph has perfect scores in all metrics except the CIDEr which has the value of zero. This is because the generated questions perfectly match the ground-truth questions, and the low CIDEr score is because this score gives high weights to keywords and this dataset has repetitive questions. The results also show a decrease in the BLEU scores between the first and the last questions. This decrease is higher in BLEU scores with larger n-grams. The METEOR and ROUGE show decreases of 50% and 11% in the second question, respectively, and stable scores in the later questions. In contrast, the CIDEr score increases from



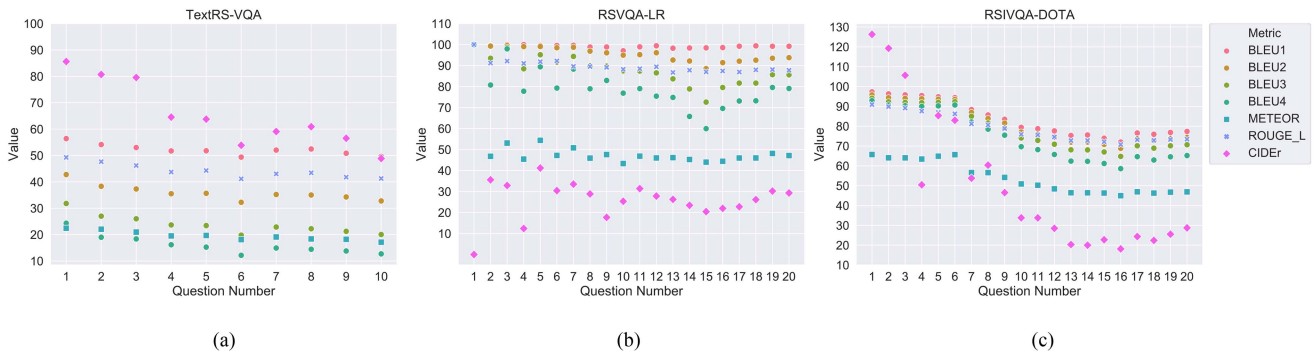


Fig. 6. Results of fine-tuning the GPT-2 model in terms of accuracy metrics for the datasets. (a) TextRS-VQA. (b) RSVQA-LR. (c) RSIVQA-DOTA.

TABLE II  
RESULTS OF FINE-TUNING THE GPT-2 MODEL IN TERMS OF ACCURACY METRICS

Dataset		BLEU1	BLEU2	BLEU3	BLEU4	ROUGE	METEOR	CIDEr
TextRS-VQA	1 <sup>st</sup> Question	56.40	42.77	31.81	24.32	49.27	22.43	85.64
	10 <sup>th</sup> Question	49.63	32.77	20.04	12.71	41.28	17.13	48.85
RSVQA-LR	1 <sup>st</sup> Question	100.0	100.0	100.0	100.0	100.0	100.0	00.00
	20 <sup>th</sup> Question	99.21	93.79	85.62	79.09	87.78	47.16	29.32
RSIVQA-DOTA	1 <sup>st</sup> Question	97.30	95.79	94.11	92.85	90.91	65.73	126.30
	20 <sup>th</sup> Question	77.34	74.31	70.67	65.20	73.46	46.89	28.74

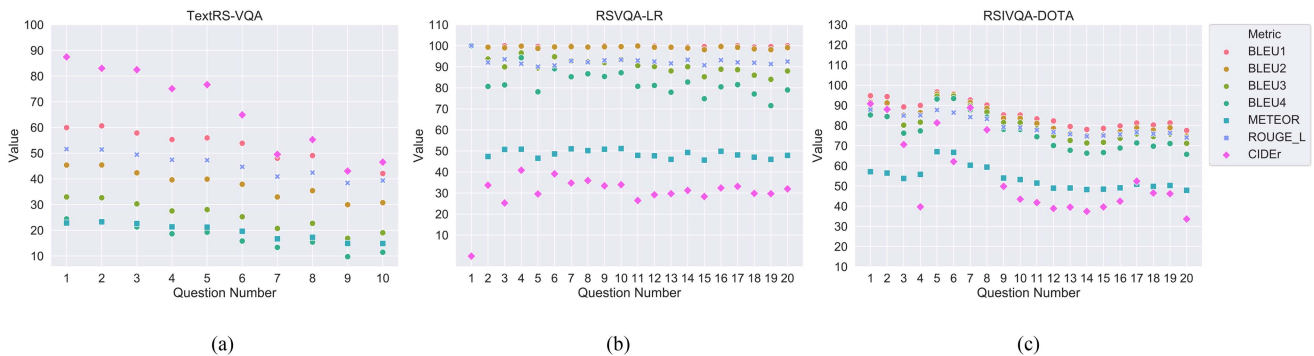


Fig. 7. Results of freezing the GPT-2 model in terms of accuracy metrics for the datasets. (a) TextRS-VQA. (b) RSVQA-LR. (c) RSIVQA-DOTA.

zero to 35% in the second question and shows stable scores in the later questions of the paragraph.

In the RSIVQA-DOTA, the scores of almost all metrics are stable for the first six questions in the paragraph. However, the scores started to drop gradually starting from question seven. This drop is rapid for the CIDEr score.

#### D. Results of Freezing the GPT-2 Model

To demonstrate the effect of fine-tuning the question generator, another experiment is conducted by freezing the weights of the language generator. Fig. 7 shows the scores of freezing the weights of the GPT-2 model, and Table III shows the scores of the first and the last questions in the generated paragraph. By comparing the results in Tables II and III, and Figs. 6 and 7. We can see that for the TextRS-VQA dataset, the scores of generating the first question are better for the frozen model. However,

the similarity scores of the question at the end of the paragraph are better for the fine-tuned model. For RSVQA-LR dataset, the results of freezing and fine-tuning the language model are identical for the first questions in the paragraph. However, the model with fixed weights shows better scores in almost all metrics as it generates more questions. For RSIVQA-DOTA, the scores of generating the first questions in the paragraph by the fine-tuned model are higher by a large margin than the fixed model. However, the scores of freezing the GPT-2 and fine-tuning it are comparable in generating later questions.

#### E. Results of Diversity Metrics

In order to provide an analysis of the diversity of the generated questions, we report in Table IV the generative strength and inventiveness of our model on the three datasets under

TABLE III  
RESULTS OF FREEZING THE GPT-2 MODEL IN TERMS OF ACCURACY METRICS

Dataset		BLEU1	BLEU2	BLEU3	BLEU4	ROUGE	METEOR	CIDEr
TextRS-VQA	1 <sup>st</sup> Question	59.93	45.37	32.96	24.40	51.59	22.83	87.45
	10 <sup>th</sup> Question	42.07	30.74	19.03	11.45	39.35	14.85	46.50
RSVQA-LR	1 <sup>st</sup> Question	100.0	100.0	100.0	100.0	100.0	100.0	00.00
	20 <sup>th</sup> Question	100.0	99.08	87.98	78.96	92.49	47.90	31.97
RSIVQA-DOTA	1 <sup>st</sup> Question	94.76	91.66	87.98	85.26	87.90	57.10	90.81
	20 <sup>th</sup> Question	77.53	74.76	71.18	65.72	74.05	47.87	33.64

TABLE IV  
RESULTS OF DIVERSITY METRICS

Dataset	Training Strategy	Strength	Inventiveness
TextRS-VQA	Fine-tuning	83.71	63.06
	Freezing	58.79	82.54
RSVQA-LR	Fine-tuning	95.95	0.007
	Freezing	45.65	09.68
RSIVQA-DOTA	Fine-tuning	81.31	0
	Freezing	62.92	15.38

different training strategies. The results show that the fine-tuned model outperforms the GPT-2 with frozen weights in terms of generative strength across all datasets. In contrast, the question generator with frozen weights has high inventiveness scores which translated as a higher capacity to generate original questions that are not present in the training set. Although our TextRS-VQA dataset has the smallest number of questions, the results demonstrate that the model’s inventiveness scores on this dataset are greater than those on other datasets. This could be attributed to the richness and the high diversity of the questions in our dataset which lead to a more diverse and expressive model. In addition, the table shows that the inventiveness scores of the fine-tuned model on the RSVQA-LR and the RSIVQA-DOTA datasets are close to zero, which is an indication of overfitting. The overfitting could be induced by the limited vocabulary size and the low diversity rate of the questions in these datasets, which make it difficult for the model to generate novel questions.

#### F. Qualitative Analysis of the Generated Questions

Fig. 8 shows examples of generated questions paragraph from one test image of each dataset along with the ground-truth questions. From the TextRS-VQA example, we can notice that the questions generated by our model contain almost all the semantic information present in the image. Although the generated questions are not perfectly matching with the ground-truth questions, some questions have the same meaning as the ground-truth questions but expressed differently, such as the reference question “is the parking space all occupied?” and the generated question “is the parking lot full of cars?”

Furthermore, our model can generate novel and valid questions that are correlated with the image content but not present in ground-truth questions such as “how many cars are in the parking lot?” We can also see some questions about objects and their attributes that are not present in the ground-truth

questions such as questions about “building” and “red car.” This proves that our model can precisely understand image content and translate it into questions. Although most of the questions in the example seem to be biased toward the presence question, and there are some incorrect questions such as “are there cars on the road empty?” but in general the questions are sensible, diverse and, in line with the semantic content of the image.

The example of the RSVQA-LR dataset shows that the generated questions are very close to the ground-truth questions, even though there are some incomplete questions such as “is the number of roads equal to the number?” In addition, all questions follow the same pattern and question types of the ground-truth questions which is in line with the quantitative results that show high similarity with the reference questions and low inventiveness scores.

Similarly, the inspection of the image and the questions generated from the sample of the RSIVQA-DOTA shows that all the generated questions are from the presence and count types, which are the only question types defined for this dataset. The generated paragraph shows some repetition in questions such as “does this picture contain container crane?”

#### G. Comparisons to the State-of-the-Art RS Captioning Methods

To demonstrate the broad applicability of our model in general text generative tasks, we compared the results of our model against state-of-the-art methods on image captioning. The experimental results are presented on two datasets. The first is UC-Merced caption dataset [7], which contains 2100 images, each image comes with five captions, and the second is the UAV dataset [8], which has 2628 images, each is annotated with three different captions. The results in Table V report the scores of the caption generated by our model on the two datasets compared to the scores of other captioning methods. Since our method is a paragraph-based approach which generates multiple captions, the results shown are for the first sentence of the paragraph. The highest results for each evaluation metric are displayed in bold, while the second-highest results are displayed in italic. The results of UC-Merced show that our model with fixed GPT-2 weights achieves the best scores in BLEU, METEOR, and CIDEr metrics, while our fine-tuned model achieves the second-best scores on the same metrics. Regarding the ROUGE metric, even though the result of our VQG model is not the best, it is still competitive with the best performance methods. The results of the UAV dataset show that our model outperforms



TABLE V  
EXPERIMENTAL RESULTS OF OUR METHOD AND OTHER STATE-OF-THE-ART METHODS ON IMAGE CAPTIONING

Dataset	Method	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE	METEOR	CIDEr
UC-Merced	VLAD+RNN [10]	63.11	51.93	46.06	42.09	58.78	29.71	200.66
	VLAD+LSTM [10]	70.16	60.85	54.96	50.30	65.20	34.64	231.31
	mRNN [6]	45.58	28.25	18.09	12.13	31.26	15.69	19.15
	mLSTM [6]	50.57	32.42	23.19	17.46	35.02	17.84	31.61
	mGRU [6]	42.56	29.99	22.91	17.98	37.97	19.41	124.82
	mGRU embedword [6]	60.94	46.24	36.80	29.81	48.20	26.14	159.54
	ConvCap [51]	70.34	56.47	46.24	38.57	59.62	28.31	190.15
	Soft-attention [10]	74.54	65.45	58.55	52.50	72.37	38.86	261.24
	Hard-attention [10]	81.57	73.12	67.02	61.82	76.98	42.63	299.47
	SD-RISC [3]	74.80	66.40	59.80	53.80	69.50	39.00	213.20
	RTRMN (semantic) [52]	55.26	45.15	39.62	35.87	55.38	25.98	180.25
	RTRMN (statistical) [52]	80.28	73.22	68.21	63.93	<b>77.26</b>	42.58	312.70
	SVM-D BOW [7]	76.35	66.64	58.69	36.54	68.01	36.54	271.42
	SVM-D CONC [7]	76.53	69.47	64.17	37.02	68.77	37.02	292.28
	<b>Our (freeze GPT-2)</b>	<b>87.44</b>	<b>81.91</b>	<b>76.93</b>	<b>72.22</b>	76.94	<b>46.31</b>	<b>330.92</b>
	<b>Our (fine-tune GPT-2)</b>	<b>85.31</b>	<b>79.40</b>	<b>74.41</b>	<b>69.69</b>	76.65	<b>45.86</b>	<b>327.68</b>
UAV	SVM-D BOW [7]	68.84	58.05	48.33	39.22	69.63	32.81	<b>391.31</b>
	SVM-D CONC [7]	65.13	56.53	48.15	39.69	69.31	32.17	389.45
	<b>Our (freeze GPT-2)</b>	<b>77.11</b>	<b>66.45</b>	<b>55.99</b>	<b>45.17</b>	<b>75.19</b>	<b>38.18</b>	<b>390.27</b>
	<b>Our (fine-tune GPT-2)</b>	<b>76.84</b>	<b>65.24</b>	<b>54.15</b>	<b>43.00</b>	<b>73.47</b>	<b>38.18</b>	<b>376.29</b>

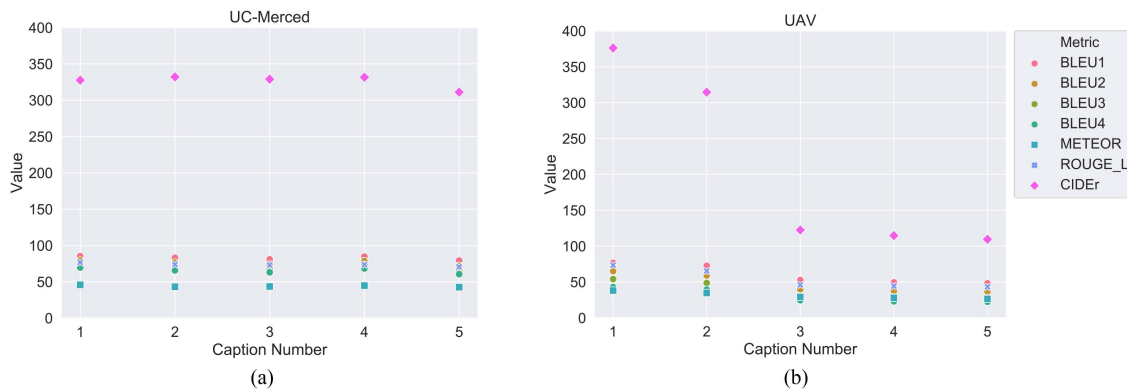


Fig. 9. Results of our method on image captioning in terms of accuracy metrics for the datasets. (a) UC-Merced. (b) UAV.

the state-of-the-art models and achieves the best results for all evaluation metrics except for the CIDEr. Furthermore, Fig. 9 shows the scores of the first five captions of the paragraph generated by our model. The figure shows that results of the captions generated for UC-Merced images are almost similar. It may attribute to the high similarity of the captions given for the UC-Merced images. In contrast, the scores of the UAV dataset are gradually decreasing as the model generates more captions in the paragraph.

#### H. Comparisons to the State-of-the-Art Computer Vision Methods on COCO Dataset

In order to verify the effectiveness of the proposed VQG model beyond the scope of RS, we compare it against several

VQG models in computer vision on the MS-COCO VQA dataset [43]. This dataset is a large-scale dataset consisting of 82 783 training images and 40 504 validation images usually used for testing. The average number of questions associated with each image is three. The comparison results are presented in Table VI. Generally, our VQG model gets the highest scores and outperforms all the existing state-of-the-art methods in terms of the four BLEU metrics. In particular, the improvement achieved by our model is significant with more than 5% in the BLEU1 score, and 6% in the BLEU2 score compared to the best model. These metrics represent the concurrence of the uni- and bi-grams between the generated and the ground-truth questions. However, the improvement is less for the higher order n-grams metrics with 1.45% improvement in the BLEU3 and 1.35% in the BLEU4 compared with Krishna et al. [33] which achieves th

TABLE VI  
EXPERIMENTAL RESULTS OF OUR METHOD AND OTHER STATE-OF-THE-ART VQG METHODS IN COMPUTER VISION ON MS-COCO-VQA DATASET

Method	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE	METEOR	CIDEr
IA2Q [53]	32.43	15.49	9.24	6.23	-	11.21	36.22
V-IA2Q [23]	36.91	17.79	10.21	6.25	-	12.39	36.39
Krishna et al. (t-space) [32]	47.40	28.95	19.93	14.49	<b>49.10</b>	18.35	85.99
Krishna et al. (z-space) [32]	50.09	32.32	24.61	16.27	-	<b>20.58</b>	<b>94.33</b>
IC2Q (Weakly supervised) [53]	30.42	13.55	6.23	4.44	-	9.42	27.42
V-IC2Q (Weakly supervised) [23]	35.40	25.55	14.94	10.78	-	13.35	42.54
Krishna et al. (Weakly supervised) [32]	31.20	16.20	11.18	6.24	40.27	12.11	35.89
I + II + CL [31]	38.94	20.30	12.47	8.10	41.27	13.47	47.32
I + II + CL + Bayes [31]	41.87	22.11	14.96	10.04	42.34	13.60	46.87
<b>Ours</b>	<b>55.34</b>	<b>38.44</b>	<b>26.06</b>	<b>17.62</b>	47.45	19.32	29.63

e second-best scores. In addition, our model achieves the second-best ROUGE and METEOR scores with 1.65%, and 1.26% less than the best model, respectively. However, the score of our model in terms of the CIDEr is lower than the state-of-the-art methods. These results demonstrate that generating rich features from the image and using an autoregressive pretrained language model like GPT-2 can significantly increase the accuracy of generating questions.

## VI. CONCLUSION

In this article, we propose a VQG model for generating questions from RS images. At first, visual features are obtained by a swin-transformer encoder which takes advantage of the self-attention mechanism to generate global and rich multiscale features from the image. Then, these features are used as tokens to guide a paragraph-based language decoder to generate multiple questions from the image. Experiments are conducted on two standard VQA datasets in addition to our proposed TextRS-VQA dataset demonstrate the effectiveness of our method in generating questions from the RS image. Furthermore, our TextRS-VQA dataset demonstrates more merit in generating richer and diverse questions from the RS image.

The current method is an image guided one that solely depends on the understanding of the image content to generate the questions. In the future, this model might be extended to depend on auxiliary data like an answer or the question's type when generating questions.

## REFERENCES

- [1] Z. Xiong, F. Zhang, Y. Wang, Y. Shi, and X. X. Zhu, "EarthNets: Empowering AI in earth observation," Dec. 2022, Accessed: Dec. 15, 2022. [Online]. Available: <http://arxiv.org/abs/2210.04936>
- [2] W. Huang, Q. Wang, and X. Li, "Denosing-based multiscale feature fusion for remote sensing image captioning," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 436–440, Mar. 2021, doi: [10.1109/LGRS.2020.2980933](https://doi.org/10.1109/LGRS.2020.2980933).
- [3] K. Fu, Y. Li, W. Zhang, H. Yu, and X. Sun, "Boosting memory with a persistent memory mechanism for remote sensing image captioning," *Remote Sens.*, vol. 12, no. 11, Jun. 2020, Art. no. 1874, doi: [10.3390/rs12111874](https://doi.org/10.3390/rs12111874).
- [4] G. Sumbul, S. Nayak, and B. Demir, "SD-RSIC: Summarization-driven deep remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6922–6934, Aug. 2021, doi: [10.1109/TGRS.2020.3031111](https://doi.org/10.1109/TGRS.2020.3031111).
- [5] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3623–3634, Jun. 2017, doi: [10.1109/TGRS.2017.2677464](https://doi.org/10.1109/TGRS.2017.2677464).
- [6] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1274–1278, Aug. 2019, doi: [10.1109/LGRS.2019.2893772](https://doi.org/10.1109/LGRS.2019.2893772).
- [7] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst.*, 2016, pp. 1–5, doi: [10.1109/CITS.2016.7546397](https://doi.org/10.1109/CITS.2016.7546397).
- [8] G. Hoxha and F. Melgani, "A novel SVM-based decoder for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 25, 2021, Art. no. 5404514, doi: [10.1109/TGRS.2021.3105004](https://doi.org/10.1109/TGRS.2021.3105004).
- [9] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word-sentence framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10532–10543, Dec. 2021, doi: [10.1109/TGRS.2020.3044054](https://doi.org/10.1109/TGRS.2020.3044054).
- [10] X. Li, X. Zhang, W. Huang, and Q. Wang, "Truncation cross entropy loss for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5246–5257, Jun. 2021, doi: [10.1109/TGRS.2020.3010106](https://doi.org/10.1109/TGRS.2020.3010106).
- [11] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018, doi: [10.1109/TGRS.2017.2776321](https://doi.org/10.1109/TGRS.2017.2776321).
- [12] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sens.*, vol. 11, no. 6, Mar. 2019, Art. no. 612, doi: [10.3390/rs11060612](https://doi.org/10.3390/rs11060612).
- [13] Y. Li, S. Fang, L. Jiao, R. Liu, and R. Shang, "A multi-level attention model for remote sensing image captions," *Remote Sens.*, vol. 12, no. 6, Mar. 2020, Art. no. 939, doi: [10.3390/rs12060939](https://doi.org/10.3390/rs12060939).
- [14] Z. Yuan, X. Li, and Q. Wang, "Exploring multi-level attention and semantic relationship for remote sensing image captioning," *IEEE Access*, vol. 8, pp. 2608–2620, 2020, doi: [10.1109/ACCESS.2019.2962195](https://doi.org/10.1109/ACCESS.2019.2962195).
- [15] Z. Zhang, W. Diao, W. Zhang, M. Yan, X. Gao, and X. Sun, "LAM: Remote sensing image captioning with label-attention mechanism," *Remote Sens.*, vol. 11, no. 20, Oct. 2019, Art. no. 2349, doi: [10.3390/rs11202349](https://doi.org/10.3390/rs11202349).
- [16] Z. Zhang, W. Zhang, W. Diao, M. Yan, X. Gao, and X. Sun, "VAA: Visual aligning attention model for remote sensing image captioning," *IEEE Access*, vol. 7, pp. 137355–137364, 2019, doi: [10.1109/ACCESS.2019.2942154](https://doi.org/10.1109/ACCESS.2019.2942154).
- [17] T. Abdullah, Y. Bazi, M. M. Al Rahhal, M. L. Mekhalif, L. Rangarajan, and M. Zuair, "TextRS: Deep bidirectional triplet network for matching text to remote sensing images," *Remote Sens.*, vol. 12, no. 3, Jan. 2020, Art. no. 405, doi: [10.3390/rs12030405](https://doi.org/10.3390/rs12030405).
- [18] 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, doi: [10.1109/TGRS.2020.2988782](https://doi.org/10.1109/TGRS.2020.2988782).

- [19] Z. Yuan, L. Mou, Z. Xiong, and X. X. Zhu, "Change detection meets visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 23, 2022, Art. no. 5630613, doi: [10.1109/TGRS.2022.3203314](https://doi.org/10.1109/TGRS.2022.3203314).
- [20] 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, 2022, Art. no. 5606514, doi: [10.1109/TGRS.2021.3079918](https://doi.org/10.1109/TGRS.2021.3079918).
- [21] S. Zhang, L. Qu, S. You, Z. Yang, and J. Zhang, "Automatic generation of grounded visual questions," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 4235–4243, doi: [10.24963/ijcai.2017/592](https://doi.org/10.24963/ijcai.2017/592).
- [22] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende, "Generating natural questions about an image," Jun. 2016, Accessed: Apr. 02, 2022. [Online]. Available: <http://arxiv.org/abs/1603.06059>
- [23] T. Scialom, P. Bordes, P.-A. Dray, J. Staiano, and P. Gallinari, "What BERT sees: Cross-modal transfer for visual question generation," Dec. 2020, Accessed: Apr. 14, 2022. [Online]. Available: <http://arxiv.org/abs/2002.10832>
- [24] U. Jain, Z. Zhang, and A. Schwing, "Creativity: Generating diverse questions using variational autoencoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5415–5424, doi: [10.1109/CVPR.2017.575](https://doi.org/10.1109/CVPR.2017.575).
- [25] A. K. Vijayakumar et al., "Diverse beam search: Decoding diverse solutions from neural sequence models," Oct. 2018, Accessed: Apr. 13, 2022. [Online]. Available: <http://arxiv.org/abs/1610.02424>
- [26] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun, "iVQA: Inverse visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8611–8619, doi: [10.1109/CVPR.2018.00898](https://doi.org/10.1109/CVPR.2018.00898).
- [27] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun, "Inverse visual question answering: A new benchmark and VQA diagnosis tool," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 460–474, Feb. 2020, doi: [10.1109/TPAMI.2018.2880185](https://doi.org/10.1109/TPAMI.2018.2880185).
- [28] Y. Alwattar and Y. Guo, "Inverse visual question answering with multi-level attentions," Dec. 2020, Accessed: Apr. 04, 2022. [Online]. Available: <http://arxiv.org/abs/1909.07583>
- [29] Y. Li et al., "Visual question generation as dual task of visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6116–6124, doi: [10.1109/CVPR.2018.00640](https://doi.org/10.1109/CVPR.2018.00640).
- [30] N. Vedd, Z. Wang, M. Rei, Y. Miao, and L. Specia, "Guiding visual question generation," Nov. 2021, Accessed: Apr. 04, 2022. [Online]. Available: <http://arxiv.org/abs/2110.08226>
- [31] Z. Fan, Z. Wei, P. Li, Y. Lan, and X. Huang, "A question type driven framework to diversify visual question generation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 4048–4054, doi: [10.24963/ijcai.2018/563](https://doi.org/10.24963/ijcai.2018/563).
- [32] S. Uppal, A. Madan, S. Bhagat, Y. Yu, and R. R. Shah, "C3VQG: Category consistent cyclic visual question generation," Jan. 2021, Accessed: Apr. 16, 2022. [Online]. Available: <http://arxiv.org/abs/2005.07771>
- [33] R. Krishna, M. Bernstein, and L. Fei-Fei, "Information maximizing visual question generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2008–2018, doi: [10.1109/CVPR.2019.00211](https://doi.org/10.1109/CVPR.2019.00211).
- [34] I. M. Mora, S. La Puente, and X. Giro-i-Nieto, "Towards automatic generation of question answer pairs from images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 1–2.
- [35] M. Sarroufi, A. Ben Abacha, and D. Demner-Fushman, "Goal-driven visual question generation from radiology images," *Information*, vol. 12, no. 8, Aug. 2021, Art. no. 334, doi: [10.3390/info12080334](https://doi.org/10.3390/info12080334).
- [36] C. Chappuis, S. Lobry, B. Kellenberger, B. L. Saux, and D. Tuia, "How to find a good image-text embedding for remote sensing visual question answering?," Sep. 2021, Accessed: May 21, 2022. [Online]. Available: <http://arxiv.org/abs/2109.11848>
- [37] T. Siebert, K. N. Clasen, M. Ravanbakhsh, and B. Demir, "Multi-modal fusion transformer for visual question answering in remote sensing," Oct. 2022, Accessed: Dec. 15, 2022. [Online]. Available: <http://arxiv.org/abs/2210.04510>
- [38] Y. Li et al., "Visual question generation as dual task of visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6116–6124.
- [39] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002, doi: [10.1109/ICCV48922.2021.00986](https://doi.org/10.1109/ICCV48922.2021.00986).
- [40] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," Oct. 2020, Accessed: Jan. 05, 2021. [Online]. Available: <http://arxiv.org/abs/2010.11929>
- [41] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," Art. no. 24.
- [42] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [43] A. Agrawal et al., "VQA: Visual question answering," Oct. 2016, Accessed: Jun. 01, 2022. [Online]. Available: <http://arxiv.org/abs/1505.00468>
- [44] 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, doi: [10.1109/TGRS.2017.2685945](https://doi.org/10.1109/TGRS.2017.2685945).
- [45] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 197–209, Nov. 2018, doi: [10.1016/j.isprsjprs.2018.01.004](https://doi.org/10.1016/j.isprsjprs.2018.01.004).
- [46] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279, doi: [10.1145/1869790.1869829](https://doi.org/10.1145/1869790.1869829).
- [47] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017, doi: [10.1109/JPROC.2017.2675998](https://doi.org/10.1109/JPROC.2017.2675998).
- [48] 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.*, 2018, pp. 3974–3983, doi: [10.1109/CVPR.2018.00418](https://doi.org/10.1109/CVPR.2018.00418).
- [49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Jan. 2017, Accessed: Jun. 02, 2022. [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [50] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2001, Art. no. 311, doi: [10.3115/1073083.1073135](https://doi.org/10.3115/1073083.1073135).
- [51] 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, doi: [10.3115/v1/W14-3348](https://doi.org/10.3115/v1/W14-3348).
- [52] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," Jun. 2015, Accessed: Jun. 02, 2022. [Online]. Available: <http://arxiv.org/abs/1411.5726>

**Laila Bashmal** (Graduate Student Member, IEEE) received the B.S. degree in computer science from the University of Dammam, Dammam, Saudi Arabia, in 2011, and the M.Sc. degree in computer engineering in 2018 from King Saud University, Riyadh, Saudi Arabia, where she is currently working toward the Ph.D. degree in computer engineering.

Her research interests include machine learning and image processing with applications to remote sensing image analysis.



**Yakoub Bazi** (Senior Member, IEEE) received the State Engineer and M.Sc. degrees in electronics from the University of Batna, Batna, Algeria, in 1994 and 2000, respectively, and the Ph.D. degree in information and communication technology from the University of Trento, Trento, Italy, in 2005.

From 2000 to 2002, he was a Lecturer with the University of M'Sila, M'Sila, Algeria. In 2006, he joined the University of Trento, as a Postdoctoral Researcher. From 2006 to 2009, he was an Assistant Professor with the College of Engineering, Al-Jouf University, Sakakah, Saudi Arabia. He is currently a Full Professor of Computer Engineering with the College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He is a referee for several international journals. His research interests include remote sensing, signal/image medical analysis, and computer vision.

Dr. Bazi is an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and the *International Journal of Remote Sensing*.



**Farid Melgani** (Fellow, IEEE) received the State Engineer degree in electronics from the University of Batna, Algeria, in 1994, the M.Sc. degree in electrical engineering from the University of Baghdad, Baghdad, Iraq, in 1999, and the Ph.D. degree in electronic and computer engineering from the University of Genoa, Genoa, Italy, in 2003.

He is currently a Full Professor of Telecommunications with the Department of Information Engineering and Computer Science, University of Trento, Trento, Italy, where he teaches pattern recognition, machine learning, and digital transmission. He is the Head of the Signal Processing and Recognition Laboratory, and the Dean of Undergrad and Grad Studies at the same department. He is coauthor of more than 260 scientific publications. His research interests are in the areas of remote sensing, signal/image processing, pattern recognition, machine learning, and computer vision.

Dr. Melgani is currently an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *International Journal of Remote Sensing*, and IEEE JOURNAL ON MINIATURIZATION FOR AIR AND SPACE SYSTEMS.



**Riccardo Ricci** (Graduate Student Member, IEEE) received the bachelor's and master's degrees in information and communication engineering in 2019 and 2021, respectively, from the University of Trento, Trento, Italy, where he is currently working toward the Ph.D. degree, focusing on the intersection of natural language processing and computer vision.

His research interests include natural language processing, computer vision, and the different ways of human-machine interaction and machine-machine interaction that those two modalities enable.



**Mohamad M. Al Rahhal** (Senior Member, IEEE) received the B.Sc. degree in computer engineering from Aleppo University, Aleppo, Syria, in 2002, the M.Sc. degree in information technology from Hamdard University, New Delhi, India, in 2005, and the Ph.D. degree in computer engineering from King Saud University, Riyadh, Saudi Arabia, in 2015.

From 2006 to 2012, he was a Lecturer with Al-Jouf University, Sakakah, Saudi Arabia. He is an Associate Professor with the College of Applied Computer Engineering, King Saud University. His research interests include signal/image medical analysis, remote sensing, and computer vision.



**Mansour Zuair** (Member, IEEE) received the B.S. degree in computer engineering from King Saud University, Riyadh, Saudi Arabia, in 1986, and the M.S. and Ph.D. degrees in computer engineering from Syracuse University, Syracuse, NY, USA, in 1989 and 1996, respectively.

He was the Chairman of the Computer Engineering Department, King Saud University, from 2003 to 2006, the Vice Dean from 2009 to 2015, and has been the Dean of the College of Computer and Information Sciences, since 2016. He is currently an Associate Professor with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University. His research interests include computer architecture, computer networks, and signal processing.