

Visual Question Answering on Multiple Remote Sensing Image Modalities

Hichem Boussaid^{1,*,\dagger}, Lucrezia Tosato^{1,2,*,\dagger}, Flora Weissgerber²
Camille Kurtz¹, Laurent Wendling¹, Sylvain Lobry^{1,\dagger}

¹LIPADE, Université Paris Cité, France

²ONERA, France

^{\dagger}corresponding authors: {hichem.boussaid, lucrezia.tosato}@etu.u-paris.fr

Abstract

The extraction of visual features is an essential step in Visual Question Answering (VQA). Building a good visual representation of the analyzed scene is indeed one of the essential keys for the system to be able to correctly understand the latter in order to answer complex questions. In many fields such as remote sensing, the visual feature extraction step could benefit significantly from leveraging different image modalities carrying complementary spectral, spatial and contextual information. In this work, we propose to add multiple image modalities to VQA in the particular context of remote sensing, leading to a novel task for the computer vision community. To this end, we introduce a new VQA dataset, named TAMMI (Text and Multi-Modal Imagery) with diverse questions on scenes described by three different modalities (very high resolution RGB, multi-spectral imaging data and synthetic aperture radar). Thanks to an automated pipeline, this dataset can be easily extended according to experimental needs. We also propose the MM-RSVQA (Multi-modal Multi-resolution Remote Sensing Visual Question Answering) model, based on VisualBERT, a vision-language transformer, to effectively combine the multiple image modalities and text through a trainable fusion process. A preliminary experimental study shows promising results of our methodology on this challenging dataset, with an accuracy of **65.56%** on the targeted VQA task. This pioneering work paves the way for the community to a new multi-modal multi-resolution VQA task that can be applied in other imaging domains (such as medical imaging) where multi-modality can enrich the visual representation of a scene. The dataset and code are available at <https://tammi.sylvainlobry.com/>.

*Hichem Boussaid and Lucrezia Tosato contributed equally.
This work is supported by Agence Nationale de la Recherche (ANR) under the ANR-21-CE23-0011 project. The experiments conducted in this study were performed using HPC/AI resources provided by GENCI-IDRIS (Grant 2023-AD011012735R2).

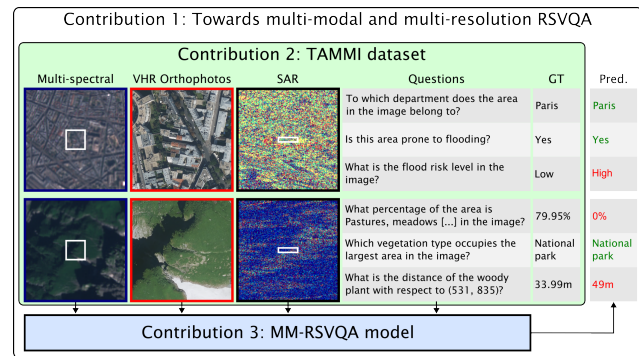


Figure 1. Summary of our contributions. We introduce a new task in the computer vision community with multi-modal and multi-resolution Visual Question Answering (VQA) on remote sensing images. We introduce a new dataset, TAMMI, associating question/answer pairs to multi-spectral, Very High-Resolution (VHR) orthophotos and Synthetic Aperture Radar (SAR) images triplets. In these examples, the white rectangle in the multi-spectral and SAR images corresponds to the extent of the VHR image. Finally, we propose a new model for this task referred as MM-RSVQA.

1. Introduction

The task of Visual Question Answering (VQA) aims at providing natural language answers to free-form, open-ended question about an image [3]. On natural images, recent advances in the computer vision and natural language processing communities have shown great improvements to standard VQA benchmarks. In particular, Large Language Models (LLM) can now be used to perform knowledge-based VQA, which requires external knowledge and commonsense reasoning [14]. These models are however more limited when used beyond natural images [16, 19].

VQA has been proposed for the medical domain [13] and is an active field of research in the medical and computer vision communities [20]. The VQA task is also used for extracting information from remote sensing data [21]. In such thematic domains, there are significant challenges compared to VQA for natural images. In particular, the lim-

ited availability of high-quality training data and the variability of information found in different imaging modalities have been active research topics [16, 20, 36].

It is commonly agreed that VQA methods for both of the remote sensing and medical imaging fields would benefit from information contained in different modalities of images [23]. Indeed, multiple image modalities are often used in these fields to obtain different, specific and complementary (spectral, spatial and contextual) information about a single scene. In remote sensing in particular, multi-modal data allows to obtain different information at multiple resolutions and from various wavelengths, from the 400nm of ultraviolet to the 5cm of radar wave. Among others, Very-High Resolution (VHR) data offers images at sub-meter resolutions. With Multi-Spectral (MS) data, it is possible to characterize different ground materials thanks to their different spectral responses. Synthetic Aperture Radar (SAR) offers imaging capacities highlighting man-made objects and giving physical information about objects [17]. While previous works have explored VQA from multi-modal images [31], using RGB data from the Sentinel-2 satellites and SAR data from Sentinel-1 on existing datasets, dedicated datasets and methods are necessary.

Our first contribution (highlighted in Figure 1) is to tackle the task of VQA from multi-modal and multi-resolution images leading to a novel challenging task for the computer vision community. Our second contribution is a new dataset named TAMMI, built from openly available data sources. This dataset combines three modalities (VHR, MS and SAR) and we make it openly available. We also share with the vision community an automated pipeline to easily extend this dataset (e.g. to new geographical areas or modalities) according to experimental needs. MS and SAR can be used dynamically at different sizes to provide different levels of context. Finally, our third contribution is a baseline model for the targeted multi-modal VQA task named MM-RSVQA. The proposed architecture is able to take as input the three image modalities, with specifically pre-trained feature extractors. To effectively integrate the multi-modal visual features, along with the textual features of the question, we use VisualBERT, a recent vision-language transformer, leading to a trainable fusion scheme. A subset of the dataset and code is available [on this link](#).

2. Related works

VQA has the objective of predicting a natural language response to an open-ended question related to an image [3]. This task bridges visual reasoning with semantics expressed in natural language, providing new challenges for the computer vision community. Models relying on attention [2, 34, 40] to extract relevant features have been proposed. Recently, foundation models such as CLIP [24] have shown remarkable success on the VQA downstream task [26]. How-

ever, such approaches do not translate directly to thematic applications of the VQA task [9].

The task of VQA for remote sensing (RSVQA) is introduced in [21]. In this work, features are extracted using a Convolution Neural Network (CNN) and a Recursive Neural Network (RNN) and fused together with a point-wise multiplication. Improvements in the feature extractor have been proposed by Felix et al. [11], inspired by LXMERT [29] enhancing the results using an object detection step implemented via the Faster-RCNN model as a visual encoder and BERT for the language part. The Fourier transform is also used in [37] to extract structural information from complex remote sensing data, thereby improving the generalizability across various domains. In [6], an object detector and a classifier are used to create a caption of the image, then used to answer the question. LLMs such as BERT [8] have been used for textual features extraction [4] or as the main component of the model [5]. The fusion step of the two representations is done in [11] with a cross-modal transformer encoder. Another attention-based method is proposed in [39], with a fusion module based on mutual attention. In [30], a method that uses segmentation maps to guide the attention is introduced. In parallel, training techniques such as self-supervised curriculum learning [35] have been shown to be efficient to obtain a common representation of textual and visual features. Improvements to the language processing have been introduced as well: in [36] an augmentation is applied to translate each question in multiple languages and then translate it back to English, improving the diversity of the formulation of questions.

Datasets Several datasets have been created for RSVQA. This problem is first considered in [21] which introduced two RSVQA datasets composed of image/question/answer triplets. The first one, called "Low Resolution (LR)", is based on images from the Sentinel-2 sensors (optical images with a spatial resolution of 10m) acquired over The Netherlands. The second one, called "High Resolution (HR)", uses RGB aerial images with a spatial resolution of 15cm extracted from the USGS High Resolution Orthoimagery database, covering urban areas of the United States. In these two datasets, the nature of the questions, as well as the distribution of answers, is highly unbalanced (e.g. '0' in the HR dataset has a frequency of 60.9% for the numerical answer). In addition, the number of different answers is very limited, with 9 possible answers for LR and 98 for HR. RSVQAxBEN [22] dataset proposes a larger number of samples and introduces new objects of interest (land cover classes) with a new form of complexity (logical formulas). The area of interest is also different, as it covers many European countries thanks to images from the Sentinel-2 satellites. However, even in this dataset, the imbalance in the distribution of answers remains. In order to increase the diversity of questions, Zheng et al. [39] have exploited five pre-

annotated datasets, three for scene classification and two for object detection for the automatic construction of questions and answers.

These datasets have certain limitations in common. First, most of them have a limited number of samples, which may reduce the ability of deep learning models to generalise effectively [21]. In addition, the diversity of questions and answers is often restricted, which can lead to biases and gaps in model performances [7]. Finally, these datasets focus on the use of RGB images, not leveraging other modalities, sensors and resolutions.

Integrating SAR imagery with natural language remains under-explored and challenging due to its unique data structure. Deep learning models often struggle with SAR-based tasks requiring precise quantification, such as target size estimation or object counting [38]. However, they are effective in tasks involving spatial relationships, like proximity identification or density assessment within SAR scenes. SAR has been used in RSVQA for tasks such as scattering pattern classification [1], ship detection and counting [32], and in combination with optical images for land cover related questions [31].

In this work, beyond sharing a new vision task with the community, we address existing gaps in the current state of the art by creating a diverse dataset that spans multiple complementary image modalities, resolutions, and geographic regions, encompassing a wide range of terrain types and incorporating various question typologies. As a first baseline on this task and dataset, we embed state of the art transformer-based fusion techniques proven effective through a new model for tackling multi-modal challenges.

3. Dataset

For the purpose of advancing RSVQA capabilities we develop TAMMI (*Text and Multi-Modal Imagery*), a large multi-modal and multi-resolution dataset. TAMMI includes images/question/answer triplets on three French departments, shown in Figure 2. The following subsections outline each phase of the dataset creation, detailing data sources and processing steps involved.

3.1. Image modalities

Very high-resolution orthophotos (BDOrtho) BDOrtho is a dataset of aerial VHR optical images acquired by the french National Geographic Institute (IGN). It provides an accurate and detailed photographic representation of the French territory at 20cm resolution. Images are provided in RGB and updated every three years. For our dataset, the most recent images are chosen for each department: images from department 74 are from 2020, and images for department 34, 75, 92, 93, 94 are from 2021. To ensure consistency in the information provided by the images, the same years are maintained for the images of the other modalities.

Multi-spectral data (Sentinel-2) Sentinel-2 is a mission of the European Union’s Copernicus program, designed to provide multi-spectral images of land. Sentinel-2 captures data in 13 bands of the electromagnetic spectrum [10], including near-infrared (NIR), visible and short-wave infrared (SWIR) with a spatial resolution of 10m, 20m, or 60m depending on the band. Sentinel-2 images are acquired every five days when both Sentinel-2A and Sentinel-2B are active. The images are distributed as Level-1C (L1C) products which provide top-of-atmosphere (TOA) reflectance, and Level-2A (L2A) that offer surface reflectance after atmospheric correction. Since our dataset includes SAR images which are not affected by atmospheric conditions and BDOrtho images which are taken during sunny days and do not need any atmospheric correction, we select L1C products. Furthermore, only images with a cloud coverage under 3% are selected.

Synthetic Aperture Radar data (Sentinel-1) The Sentinel-1 satellite is also part of the European Union’s Copernicus program, designed to provide full coverage of Europe every 6 days. Sentinel-1 acquires SAR images by sending radar pulses towards the earth’s surface and measuring the backscattered signal. Since the radar pulses propagate through clouds, the acquisition rate is independent of the atmospheric conditions. In our dataset, we use Interferometric Wide (IW) swath mode. IW is the main acquisition mode over land and has a resolution of 5m in range (across satellite trajectory) and 20m in azimuth (along trajectory). In this acquisition mode, the swath is divided in three sub-swathes. The satellite acquires tiles, called *burst*, in each sub-swath sequentially with an overlap between the bursts. The bursts are processed as separate Single Look Complex (SLC) images. In the Sentinel-1 SLC products, sequentially acquired bursts of the same sub-swath are included into a single image separated by black bands. Sentinel-1 acquires data in two polarization channels (VV and VH), providing additional information about the characteristics of materials on the planet surface. In this work, we use the amplitude of the SLC level 1 images for both polarization channels.

3.2. Image triplets

To build the question/answer pairs, we select the geographical extent of the VHR patches p_{VHR} . The VHR patches are extracted from the division of the VHR tiles (of dimension 25000×25000 pixels) of our areas of interest (shown in Figure 2) into patches of size 1000×1000 pixels (i.e. $200m \times 200m$). The areas of interest are selected for their diverse landscapes. Specifically, the Île-de-France region (departments 75, 92, 93, 94) is predominantly urban, while the Haute-Savoie region (department 74) features mountainous terrain, and the Hérault region (department 34) is maritime.

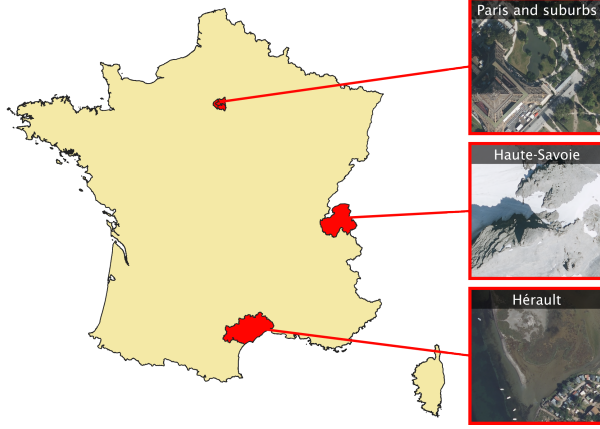


Figure 2. Geographical extent of the TAMMI dataset, covering selected regions (highlighted in red) in Metropolitan France: Paris and inner suburbs (departments 75, 92, 93, 94), an urban region; Haute-Savoie (74), a mountainous region; and Hérault (34), a sea-side region. For each of these regions, we show a VHR sample from the dataset.

Since MS (p_{MS}) and SAR (p_{SAR}) patches are designed to offer spatial context to the model, their extent (respectively L_{MS} and L_{SAR}) is defined by the user and the patches are cropped at execution time. To align the images of all three modalities, we retrieve the latitude, longitude and altitude of the central pixel for each VHR patch. We then associate a MS and a SAR tile centered on the position of the central pixel of the VHR patch. The MS tiles are geocoded. Therefore, the latitude and longitude are converted to the image pixel space. For SAR images, we use the method of [33] to project the position of the central pixel of p_{VHR} . To produce continuous SAR patches, the SAR images are debursted (removing black lines and burst overlaps), and pre-processed with tail-value removal based on histogram analysis for each polarization, then a transformation to decibels and a normalization are applied. The pre-processing steps are detailed in the supplementary materials.

3.3. Question/answer pairs

Inspired by [15], we propose an automatic approach for the construction of question/answer pairs associated to each geographical area covered by the VHR patch, using:

- **BDTopo** provided by IGN (French Geographical Institute). It is an official vectorial description of the French territory, including generic geographical objects (e.g. buildings and water areas), and specific ones (e.g. museums and lakes);
- **Flooding risks database (TRI)** from *Géorisques*. This dataset corresponds to the areas at significant risk of flooding, including management zoning, floodable area zoning, water level zoning, etc.;
- **Urban units 2020 (BU20)** of INSEE (French National In-

stitute of Statistics and Economic Studies). This database corresponds to the urban zoning of French cities according to the continuity of the constructions and the number of inhabitants;

- **CORINE Land Cover (CLC) 2018** produced as part of the Copernicus European land monitoring program and managed by the European environment agency. It is a database of land cover and land use classifications, on three different levels of hierarchy with different levels of details. In this work, we consider the Level-3 having the most detailed classes (44);
- **European Mountain Areas (EMA)** provided by the European environment agency. It contains information about the geometry, the area and the name of the mountains in Europe.

Construction of questions For a given VHR patch p_{VHR} , we retrieve the collection of geo-located objects \mathbf{o} that are present in $E_{p_{VHR}}$, the geographical extent of p_{VHR} . A collection of objects $\mathbf{o}_C = \{\mathbf{o}_i^C\}$ are characterised by a class C . A class is one element present in BDTopo, TRI, CLC, EMA, or an Urban Unit. We define five question types that can be divided into 21 sub-questions:

1. **Presence Questions:** questions about the existence of certain objects or features in the image.
 - (a) **Presence:** answered with *yes* if the cardinality of the set of object from the class $|\mathbf{o}_C| > 0$ and *no* otherwise, where $C \in \text{BDTopo}$, e.g. "Is there a road in the image?"
 - (b) **Mountain presence** answered with *yes* if the cardinality of the set of object from EMA $|\mathbf{o}_C| > 0$ and *no* otherwise, e.g. "Are there any mountains in the image?"
 - (c) **Flood presence** answered with *yes* if the cardinality of the set of objects from TRI $|\mathbf{o}_C| > 0$ and *no* otherwise, e.g. "Is this area prone to flooding?"
2. **Quantity Questions:** questions about the number or amount of certain objects or features in the image.
 - (a) **Count:** answered by $|\mathbf{o}_C|$, where $C \in \text{BDTopo}$, e.g. "How many buildings are in the image?"
 - (b) **Density** answered by $\frac{(\sum_{i=1}^{|\mathbf{o}_C|} a(\mathbf{o}_i^C)) \times 100}{a(p_{VHR})}$ where $a(\cdot)$ is a function returning the area, and $C \in \text{BDTopo}$, e.g. "What is the religious buildings density?"
 - (c) **Area:** answered by $a(\mathbf{o}_i^C)$, where $C \in \text{BDTopo}$, e.g. "What is the area of the lake?"
 - (d) **Percentage:** answered by $\frac{(\sum_{i=1}^{|\mathbf{o}_C|} a(\mathbf{o}_i^C)) \times 100}{a(p_{VHR})}$ where $C \in \text{CLC}$, e.g. "What percentage of the area is wetland?"
3. **Location Questions:** questions about the location of certain objects or features in the image.
 - (a) **Absolute location:** answered by $\text{loc}(\mathbf{o}_i^C)$ where $\text{loc}(\cdot)$ is a function returning the position (both as

the exact coordinate in the image space, and as the position in a 3×3 grid dividing the image) of a specific object, and $C \in \text{BDTopo}$,

e.g. "Where is the largest vegetation area?"

4. **Classification Questions:** questions about the classification or type of certain objects or features in the image.

(a) **Water bodies:** answered by the class C of $\max a(\mathbf{o}^C)$ or $\min a(\mathbf{o}^C)$, where $C \in \text{BDTopo}$'s water category (see appendix),

e.g. "What type of water body occupies the largest area in the image?"

(b) **Vegetation zones** answered by the class C of $\max a(\mathbf{o}^C)$ or $\min a(\mathbf{o}^C)$, where $C \in \text{BDTopo}$'s vegetation category (see appendix),

e.g. "Which vegetation type occupies the smallest area in the image?"

(c) **Mountain range name:** asked only if the answer of mountain presence is *yes*, and answered by $n(\mathbf{o}_i)$ where $n(\cdot)$ is a function returning the name of the mountain range present in the geographical extent of p_{VHR} ,

e.g. "What is the name of the mountain range in the image?"

(d) **Flood level:** answered by $\text{flood}(Ep_{\text{VHR}})$ where $\text{flood}(\cdot)$ is a function returning the highest flood risk present in p_{VHR} . The possible answers of this function are 'low', 'medium', and 'high'.

e.g. "What is the flood risk level?"

(e) **Flood type:** answered by $\text{type}(Ep_{\text{VHR}})$ where $\text{type}(\cdot)$ is a function returning all types of flood risks present in p_{VHR} . The possible answers of this function are 'River overflows', 'Runoff', 'Sea Flooding' and 'GroundWater overflows'

e.g. "What is the nature of the flood risk in this area?"

(f) **Land cover:** answered by the CLC class with the largest (or smallest) area,

e.g. "Which land cover category occupies the largest area in the image?"

(g) **Urban** : answered by $\text{urban}(Ep_{\text{VHR}})$ where $\text{urban}(\cdot)$ is a function returning the urban classification (City Center, Suburb, Isolated City, Outside Urban Unit) based on BU20,

e.g. "What is the urban classification of the area in the image?"

(h) **Department:** answered by $\text{dpt}(Ep_{\text{VHR}})$ where $\text{dpt}(\cdot)$ is a function returning the name of the department of Ep_{VHR} ,

e.g. "To which department does the area in the image belong to?"

(i) **Region:** answered by $\text{reg}(Ep_{\text{VHR}})$ where $\text{reg}(\cdot)$ is a function returning the name of the region of Ep_{VHR} ,

e.g. "To which region does the area in the image

belong to?"

5. **Relational Analysis Questions:** questions seeking to understand the relationships, comparisons, and distances between various objects or features in the image (note that for these questions $C \in \text{BDTopo}$).

(a) **Distance:** answered by $d(\mathbf{o}_i^{C_1}, \mathbf{o}_j^{C_2})$ where $d(\cdot, \cdot)$ is a function returning the distance (in meters) between two objects (note that $\mathbf{o}_i^{C_1}$ can be replaced by a position pos),

e.g. "What is the distance between the museum and the religious place?"

(b) **Comparison:** answered with *yes* if $|\mathbf{o}_{C_1}| > |\mathbf{o}_{C_2}|$, *no* otherwise,

e.g. "Are there more buildings than roads?"

(c) **Relative location:** answered by $l^{\mathbf{o}_i^{C_1}}(\mathbf{o}_j^{C_2})$ where $l^{\mathbf{o}_i^{C_1}}(\cdot)$ is a function returning the position (both as the exact coordinate in the image space, and as the position in a slice of a regular octagon) of an object with respect to $\mathbf{o}_i^{C_1}$,

e.g. "What is the relative position of the monument with respect to the river?"

(d) **Nearest:** answered by $l_{\text{pos}}(\mathbf{o}_C)$ where $l_{\text{pos}}(\cdot)$ is a function returning the position (both as the exact coordinate in the image space, and as the position in a 3×3 grid dividing the image) of the nearest element of a collection from the position pos ,

e.g. "Where is the closest road to (142, 221)?"

Balancing the dataset One of the challenges of constructing a VQA dataset stochastically is to balance both question types and the answer types to reduce language biases [12]. In this work, we perform the balancing jointly at the dataset and department levels.

We perform the questions/answers pairs construction at the department level. For each VHR patch p_{VHR} 10 questions per type are randomly constructed. We iterate through the questions created for each patch, and select questions based on four goals: 1) having an equal number of questions for each type; 2) having a number of question proportional to the number of VHR patches per department; 3) having a balanced distribution of answers for each type of question; and 4) not having more than 50 question/answer pairs for each patch.

In the case of question types with a fixed number of possible answers N_A such as Presence (e.g. Yes/No, $N_A = 2$), the number of questions per answer is defined as $N_{Q,A} = \frac{N_P}{\min(N_A, 10)}$, where N_P is the number of patches in the department. For Location questions, the locations are binned in the 3×3 square grid, leading to $N_A = 9$. For questions with free numerical answers such as Area or Count, the number of each of the numerical value is capped to $N_{Q,A} = \frac{N_P}{\log(x+3)}$, where x is the numerical answer and

Name	Modalities	Resolution	Annotation	# Images	# Questions	# Q. Types	# Unique A.
TAMMI	BDOrtho	0.2m	BDTopo,	282'852	3'162'514	21	109'737
	S2 - MS 10 channels S1 - VV/VH/Ratio channels	10-20m 5 × 20m	TRI, BU20, CLC, EMA				
RSVQAxBEN [22]	S2 - RGB	10m	BigEarthNet [27]	590'326	14'758'150	2	26'875
RSIVQA [39]	Various RGB	0.15–8m	Manual/Other	37'264	111'134	4	579
RSVQA LR [21]	S2 - RGB	10m	OSM	772	77'232	4	9
RSVQA HR [21]	USGS Ortho	0.15m	OSM	10'659	1'066'316	4	55

Table 1. Summary of existing RSVQA datasets, showing modalities (S1: Sentinel-1, S2: Sentinel-2), spatial resolutions, annotation sources, and dataset sizes by image and question counts.

$x + 3$ is chosen empirically. This allows us to control the distribution of each possible answer.

Note that there is an exception for Department (3.b) and Region (3.c) questions. For each department, the answer will always be the same. To overcome this, each possible answer is capped to 2'473 for departments (the number of occurrences of the answer Paris) and 16'274 for regions (the number of occurrences of Île-de-France). This ensures a balanced distribution of the answers on the dataset level. An overview on TAMMI and other RSVQA datasets is provided in Table 1. It shows that the proposed dataset significantly improves on the number of question types and in diversity of answers compared to existing RSVQA datasets.

4. Method

We propose a new methodology to tackle the VQA task in a multi-modal image setting. The general outline of the proposed architecture, named MM-RSVQA (*Multi-modal Multi-resolution RSVQA*), is presented in Figure 3. We first describe the feature extraction process from the different modalities (multi-modal images and questions) in subsection 4.1. We then present in subsection 4.2 the data fusion part of our model. Finally, the prediction of the answers is described in subsection 4.3.

4.1. Multi-modal features extraction

As discussed in section 3, the questions are based on the geographical extent of the VHR patch p_{VHR} of size 1000×1000 pixels. Our objective is to extract relevant visual features from the p_{VHR} patch, the MS patch p_{MS} of spatial size $L_{MS} \times L_{MS}$ and the SAR one p_{SAR} of spatial size $L_{SAR} \times L_{SAR}$ to obtain useful characteristics for the VQA task. The size of the patches p_{MS} and p_{SAR} is a hyper-parameter that allows taking more or less context from the MS and SAR data into account. In this work, we consider the 10 bands with 10m and 20m resolution from the MS patches, and the SAR patch is the VV, VH, and the ratio VV/VH channels [31]. Each of the p_{VHR} , p_{MS} and p_{SAR} patches is passed through a separate feature extractor.

The fourth modality is the textual data corresponding to the question. The question is tokenized using DistilBERT [25], and the tokens are converted into embeddings.

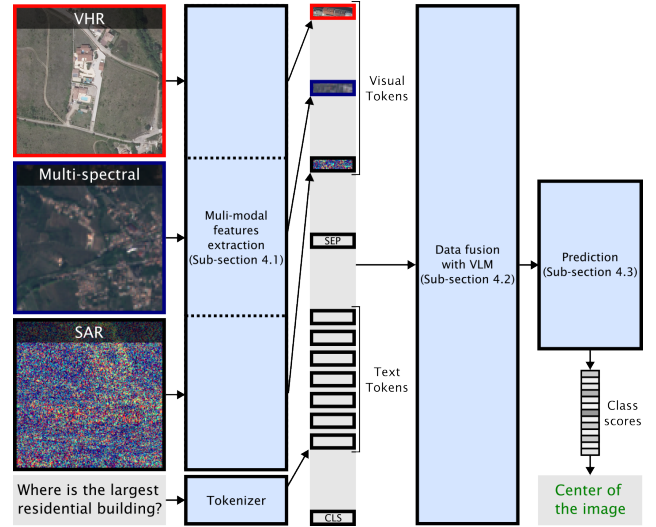


Figure 3. Graphical outline of the proposed MM-RSVQA (*Multi-modal Multi-resolution RSVQA*) architecture. The inputs of the model (multi-modal imagery and textual question) are represented on the left and the output (predicted answer) is on the bottom right. First, we extract features from each image modality and perform a text embedding. These features are passed through a vision-language model (VLM) to obtain a vector which can be classified among a set of pre-defined answers. The different blocks composing the system are detailed in section 4.

4.2. Transformer-based data fusion

We first project the visual features obtained through the feature extractor to a 768-dimensional vector through a linear layer. The 768-dimensional visual feature is concatenated with a [SEP] token, which marks the separation between visual and textual tokens, and then with the textual embeddings. Finally, a [cls] token is added to represent the entire input.

To process the diverse features, we leverage VisualBERT [18], a state-of-the-art vision-language model designed for the integration of visual and textual features. VisualBERT uses a stack of transformer layers to align regions in the images with the text input through self-attention mechanisms. Pre-trained on image-caption pairs, VisualBERT uses objectives as image-text matching and masked

language modeling. It is effective for various tasks, such as VQA, visual commonsense reasoning, natural language for visual reasoning, and region-to-phrase grounding.

The cornerstone of the proposed method is therefore to learn jointly and in an end-to-end manner an optimized representation of the data and a fusion process. The latter makes it possible to exploit the complementarity of multi-modal information and to weight the best modalities, according to the visual content of the scene studied, the types of questions, and the expected answers.

4.3. Prediction

The data fusion part of our architecture outputs a 768-dimensional vector that represents the fused multi-modal features. In this model, we frame the VQA as a classification task. Therefore, this vector is passed through a linear layer that maps it to a k -dimensional output.

5. Results and discussion

We present preliminary performances of the proposed method MM-RSVQA on the TAMMI dataset. With respect to the RSVQA task, we discuss the contribution (and complementarity) of the different imaging modalities with regard to the different types of questions (and answers) constituting the dataset.

5.1. Experimental settings

Dataset splitting The dataset is randomly split into training, validation, and test sets based on the images with a proportion of 60%, 20%, and 20% respectively. We use the validation set for the tuning of the hyper-parameters described in the rest of this section. For the training, we consider only the questions having an answer in the top $k = 1'000$ most frequent answers. This parameter is fixed for dimensionality reduction, as done in [3], [21], covering 86.6% of train answers. While we consider all the samples of the test set.

Full pipeline We use a frozen ResNet-152 model pre-trained on ImageNet for the orthophotos feature extractor and two ResNet-50 models pre-trained on BigEarthNet [28] for the MS and SAR feature extractors. We set $L_{MS} = 100$ and $L_{SAR} = 200$. We use a cross-entropy loss, optimized with Adam. For the training of MM-RSVQA, we set the learning rate to 3×10^{-5} , the batch size to 80 samples and the number of epochs to five.

5.2. Metrics

Three metrics are used to evaluate the VQA results: the per-class accuracy, overall accuracy (OA) and average accuracy (AA). The per-class accuracy is defined as the ratio of correct answer with the total number of questions for one of the 21 question types. The overall accuracy is the ratio of

correct answers with the total number of questions in the dataset. Finally, the average accuracy is the arithmetic mean of the per-class accuracies.

5.3. Quantitative results

To evaluate the proposed architecture and the TAMMI dataset, we conduct experiments using our main model MM-RSVQA and we perform ablation studies with various combinations of modalities. The results are presented in Table 2. We show the performances of the MM-RSVQA model (using VHR, MS and SAR modalities) and ablation studies across question types, highlighting the contributions of each modality in the context of this challenging task.

MM-RSVQA The proposed multi-modal multi-resolution baseline model, as seen in Table 2, achieves strong performance across most question types and outperforms models using only one or two modalities. This highlights the advantage of integrating multiple modalities for the RSVQA task. Notably, MM-RSVQA demonstrates high accuracy through categories such as presence, comparison, water and region. These results indicate that combining VHR, MS, and SAR data helps the model to better identify objects, to assess quantities, and to improve classification accuracy. For some question types, such as absolute location, area and relative location, the model shows lower accuracy. This suggests that these questions types are more challenging due to the difficulty of understanding spatial relationships or precisely locating objects. One hypothesis is that the added trainable parameters when adding MS and SAR modalities requires more training samples than for other models.

Ablation study To evaluate the contribution of the different modalities we perform ablations studies at the input level. The results, shown in Table 2 present the accuracy of different configurations: VHR only; MS RGB + VHR (only keeping the RGB bands of Sentinel-2 and VHR patch); MS + VHR (10 bands of Sentinel-2 and VHR patch); SAR + VHR; and MS + SAR (Sentinel-1 SAR images + 10 bands of Sentinel-2).

From these results we observe that using VHR only does not allow to obtain good results. Despite the fact that the questions only concern the geographical extent of the VHR patch, it appears that providing additional context, either through MS or SAR, brings better performances. This is clearly visible in classification questions such as department or urban. This suggests that the integration of SAR and multi-spectral data enhances the ability of the model to understand and classify complex features. This validate the main hypothesis of this work: additional context, even if the question is restricted to a set geographical extent, improve performances.

Category	Question Type	Ablation studies					
		MM-RSVQA	VHR	MS RGB + VHR	MS + VHR	SAR + VHR	MS + SAR
Presence Questions	Presence	96.87	95.41	96.58	96.79	97.00	96.84
	Flood Presence	98.04	94.02	97.78	97.84	96.85	98.11
	Mountain Presence	98.80	95.76	98.75	98.49	97.75	98.59
Quantity Questions	Count	51.76	49.73	50.99	50.89	50.59	51.22
	Density	26.46	26.17	26.40	26.45	26.42	26.44
	Area	10.28	10.28	10.30	10.12	10.33	10.32
	Percentage	41.66	41.58	41.66	41.60	41.66	41.62
Location Questions	Absolute Location	17.17	17.15	17.21	16.89	17.04	17.18
	Water	89.38	81.10	87.73	88.65	83.14	88.23
Classification Questions	Vegetation	53.06	48.57	50.92	51.00	48.57	50.83
	Flood Level	80.92	75.68	79.19	79.27	75.68	79.54
	Flood Type	97.91	90.60	95.68	96.93	95.54	97.32
	Land Cover	74.76	68.79	72.53	72.42	74.89	74.27
	Urban	93.78	69.20	90.33	91.58	92.40	93.32
	Department	98.96	89.78	97.39	97.58	98.03	98.47
	Region	99.98	99.96	99.99	99.92	99.99	99.99
	Mountain Name	99.89	99.94	99.96	99.91	99.89	99.96
Relation Questions	Distance	9.08	9.08	9.07	9.07	9.05	9.08
	Comparison	98.43	98.25	98.34	98.36	98.43	98.41
	Relative Location	17.99	18.10	17.77	17.62	18.17	17.74
	Nearest	21.57	21.71	20.56	20.89	20.96	20.74
	Average Accuracy	65.56	61.94	64.72	64.87	64.40	65.15
Overall Accuracy	55.11	52.22	54.41	54.49	54.29	54.70	

Table 2. Accuracy for the VQA task on the TAMMI dataset. Comparison of MM-RSVQA, VHR, MS RGB + VHR, MS + VHR, SAR + VHR, and MS + SAR across question types. Highest scores (per question type) are shown in bold.

By comparing MS RGB + VHR and MS + VHR, we can see that the performances are similar, with a small improvement for the model considering the 10 spectral bands of Sentinel-2 despite the added parameters. This validates the approach taken in other datasets considering Sentinel-2 data (RSVQA LR, RSVQAxBEN, see Table 1) which only considered the RGB channels. However, to the best of our knowledge, it is the first time that this hypothesis is experimentally demonstrated for VQA.

Regarding the SAR modality, we can see that SAR + VHR obtains similar performances to MS + VHR. This indicates that the benefits of adding context holds whether the modality. However, it appears that SAR is particularly efficient at discriminating certain features, such as land cover. Finally, our experiment with MS + SAR modalities indicates that jointly considering both modalities is a strong advantage for the model. Indeed, despite not having the very high resolution data, this model obtains the second best overall performances behind MM-RSVQA.

6. Conclusion and limitations

In this work, we address the VQA on Remote Sensing data problem by utilizing multi-modal and multi-resolution data. With these diverse modalities, we show that a model can benefit from their complementarity, in terms of coverage, spectral resolution, spatial resolution and physical information. This represents a new challenge for the vision community, as the interaction between different im-

age modalities, beyond RGB, and text remains under-explored. To support this new problem, we propose a new dataset, named TAMMI, which contains 21 types of questions based on very high resolution patches (orthophotos), high-resolution multispectral data (Sentinel-2), and Synthetic Aperture Radar (Sentinel-1) images. The dataset spans three French departments, each with distinct landscapes, allowing for diverse and challenging questions.

We introduce a baseline model to process multi-modal images and textual input data based on the VisualBERT architecture. Our results indicate that multi-resolution and multi-modal data enhance model performance. The VHR modality is useful for tasks requiring high-detail images. For tasks requiring a broader context, the contributions from the SAR and MS modalities prove to be substantial. In our pipeline, we use all three modalities together. However, ablation studies show that even when one modality is missing, the proposed model obtains better performances compared to VHR RGB orthophotos alone. This is a strong result, as this modality is the one commonly used in RSVQA.

Future work includes developing better models using specialized feature extractors for multispectral and SAR data. This dataset introduces a challenging task and, thanks to dynamic context selection, supports a variety of multi-modal applications beyond VQA. It is also easily extensible to cover more regions, promoting better generalization across diverse landscapes and environments.

References

- [1] Hossein Aghababaei and Alfred Stein. Visual Question Answering for Wishart H-Alpha Classification of Polarimetric SAR Images. In *IGARSS*, pages 11231–11234. IEEE, 2024. 3
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR Proceedings*, pages 6077–6086, 2018. 2
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV Proceedings*, pages 2425–2433, 2015. 1, 2, 7
- [4] Christel Chappuis, Vincent Mendez, Eliot Walt, Sylvain Lobry, Bertrand Le Saux, and Devis Tuia. Language transformers for remote sensing visual question answering. In *IGARSS*, pages 4855–4858. IEEE, 2022. 2
- [5] Christel Chappuis, Valérie Zermatten, Sylvain Lobry, Bertrand Le Saux, and Devis Tuia. Prompt-RSVQA: Prompting visual context to a language model for remote sensing visual question answering. In *CVPR*, pages 1372–1381, 2022. 2
- [6] Christel Chappuis, Charlotte Sertic, Nicola Santacrose, Javiera Castillo Navarro, Sylvain Lobry, Bertrand Le Saux, and Devis Tuia. Multi-task prompt-RSVQA to explicitly count objects on aerial images. In *BMVC Workshop*, 2023. 2
- [7] Christel Chappuis, Eliot Walt, Vincent Mendez, Sylvain Lobry, Bertrand Le Saux, and Devis Tuia. The curse of language biases in remote sensing VQA: the role of spatial attributes, language diversity, and the need for clear evaluation. *arXiv preprint arXiv:2311.16782*, 2023. 3
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2
- [9] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain? In *EACL*, pages 1151–1163, 2023. 2
- [10] European Space Agency. S2 mission [https://sentiwiki.copernicus.eu/web/s2-mission#S2Mission - RadiometricPerformances2 - Mission - Radiometric - Performancetrue](https://sentiwiki.copernicus.eu/web/s2-mission#S2Mission-RadiometricPerformances2-Mission-Radiometric-Performancetrue). (Accessed 14/11/2024). 3
- [11] Rafael Felix, Boris Repasky, Samuel Hodge, Reza Zolfaghari, Ehsan Abbasnejad, and Jamie Sherrah. Cross-modal visual question answering for remote sensing data. In *DICTA*, pages 1–9. IEEE, 2021. 2
- [12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR Proceedings*, pages 6904–6913, 2017. 5
- [13] Sadid A Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Henning Müller, and Matthew Lungren. Overview of image CLEF 2018 medical domain visual question answering task. *Proceeding of CLEF*, 2018. 1
- [14] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. PromptCap: Prompt-guided image captioning for SAR with GPT-3. In *ICCV Proceedings*, pages 2963–2975, 2023. 1
- [15] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR Proceedings*, pages 2901–2910, 2017. 4
- [16] Zaid Khan, Vijay Kumar BG, Samuel Schuler, Xiang Yu, Yun Fu, and Manmohan Chandraker. Q: How to specialize large vision-language models to data-scarce VQA tasks? a: Self-train on unlabeled images! In *CVPR Proceedings*, pages 15005–15015, 2023. 1, 2
- [17] Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, and Jocelyn Chanussot. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *JAG*, 112:102926, 2022. 2
- [18] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 6
- [19] Yingshu Li, Yunyi Liu, Zhanyu Wang, Xinyu Liang, Lingqiao Liu, Lei Wang, Leyang Cui, Zhaopeng Tu, Longyue Wang, and Luping Zhou. A comprehensive study of GPT-4V’s multimodal capabilities in medical imaging. *medRxiv*, pages 2023–11, 2023. 1
- [20] Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, page 102611, 2023. 1, 2
- [21] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. RSVQA: Visual question answering for remote sensing data. *TGRS*, 58(12):8555–8566, 2020. 1, 2, 3, 6, 7
- [22] Sylvain Lobry, Begüm Demir, and Devis Tuia. RSVQA meets BigEarthNet: a new, large-scale, visual question answering dataset for remote sensing. In *IGARSS*, pages 1218–1221. IEEE, 2021. 2, 6
- [23] Claudio Persello, Jan Dirk Wegner, Ronny Hänsch, Devis Tuia, Pedram Ghamisi, Mila Koeva, and Gustau Camps-Valls. Deep learning and earth observation to support the sustainable development goals: Current approaches, open challenges, and future opportunities. *GRS*, 10(2):172–200, 2022. 2
- [24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2
- [25] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 6
- [26] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt

- Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 2
- [27] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. BigEarthNet: A Large-Scale Benchmark Archive For Remote Sensing Image Understanding. In *IGARSS*, pages 5901–5904. IEEE, 2019. 6
- [28] Gencer Sumbul, Arne De Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begüm Demir, and Volker Markl. BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *GRS*, 9(3):174–180, 2021. 7
- [29] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*. Association for Computational Linguistics, 2019. 2
- [30] Lucrezia Tosato, Hichem Boussaid, Flora Weissgerber, Camille Kurtz, Laurent Wendling, and Sylvain Lobry. Segmentation-guided attention for visual question answering from remote sensing images. In *IGARSS*, pages 2750–2754. IEEE, 2024. 2
- [31] Lucrezia Tosato, Sylvain Lobry, Flora Weissgerber, and Laurent Wendling. Can SAR improve RSVQA performance? In *EUSAR*, pages 1287–1292. VDE, 2024. 2, 3, 6
- [32] Fei Wang, Chengcheng Chen, Hongyu Chen, Yugang Chang, and Weiming Zeng. A visual question answering method for SAR ship: Breaking the requirement for multimodal dataset construction and model fine-tuning. *arXiv preprint arXiv:2411.01445*, 2024. 3
- [33] Flora Weissgerber, Laurane Charrier, Cyril Thomas, Jean-Marie Nicolas, and Emmanuel Trouvé. LabSAR, a one-GCP coregistration tool for SAR–InSAR local analysis in high-mountain regions. *Frontiers in Remote Sensing*, 3, 2022. 4
- [34] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR Proceedings*, pages 21–29, 2016. 2
- [35] Zhenghang Yuan, Lichao Mou, and Xiao Xiang Zhu. Self-paced curriculum learning for visual question answering on remote sensing data. In *IGARSS*, pages 2999–3002. IEEE, 2021. 2
- [36] Zhenghang Yuan, Lichao Mou, and Xiao Xiang Zhu. Multilingual augmentation for robust visual question answering in remote sensing images. In *JURSE*, pages 1–4. IEEE, 2023. 2
- [37] Enyuan Zhao, Ziyi Wan, Xinyue Liang, Min Ye, Jie Nie, Lei Huang, et al. Frequency domain transfer learning for remote sensing visual question answering. 2
- [38] Kai Zhao and Wei Xiong. Exploring data and models in SAR ship image captioning. *IEEE Access*, 10:pp. 91150–91159, 2022. 3
- [39] Xiangtao Zheng, Binqiang Wang, Xingqian Du, and Xiaoqiang Lu. Mutual attention inception network for remote sensing visual question answering. *TGRS*, 60:1–14, 2021. 2, 6
- [40] Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji. TRAR: Routing the attention spans in transformer for visual question answering. In *ICCV Proceedings*, pages 2074–2084, 2021. 2